Origami and Geometric Constructions

By Robert J. Lang

Introduction

Compass-and-straightedge geometric constructions are familiar to most students from high-school geometry. Nowadays, they are viewed by most as a quaint curiosity of no more than academic interest. To the ancient Greeks and Egyptians, however, geometric constructions were useful tools, and for some, everyday tools, used for construction and surveying, among other activities.

The classical rules of compass-and-straightedge allow a single compass to strike arcs and transfer distances, and a single unmarked straightedge to draw straight lines; the two may not be used in combination (for example, holding the compass against the straightedge to effectively mark the latter). However, there are many variations on the general theme of geometric constructions that include use of marked rules and tools other than compasses for the construction of geometric figures.

One of the more interesting variations is the use of a folded sheet of paper for geometric construction. Like compass-and-straightedge constructions, folded-paper constructions are both academically interesting and practically useful—particularly within *origami*, the art of folding uncut sheets of paper into interesting and beautiful shapes. Modern origami design has shown that it is possible to fold shapes of unbelievable complexity, realism, and beauty from a single uncut square. Origami figures posses an aesthetic beauty that appeals to both the mathematician and the layman. Part of their appeal is the simplicity of the concept: from the simplest of beginnings springs an object of depth, subtlety, and complexity that often can be constructed by a precisely defined sequence of folding steps. However, many origami designs—even quite simple ones—require that one create the initial folds at particular locations on the square: dividing it into thirds or twelfths, for example. While one could always measure and mark these points, there is an aesthetic appeal to creating these key points, known as reference points, purely by folding.

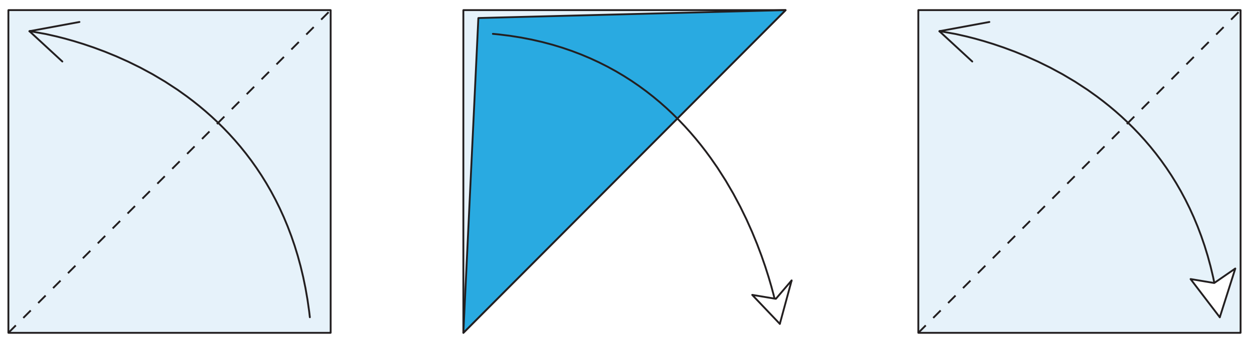
Thus, within origami, there is a practical interest in devising folding sequences for particular proportions that overlaps with the mathematical field of geometric constructions. Within this article, I will present a variety of techniques for origami geometric constructions. The field is rich and varied, with surprising connections to other branches of mathematics. I will show origami constructions based on binary divisions, and then show how these can be extended construction of proportions that are arbitrary rational fractions. Certain irrational proportions are also constructible with origami; I will present several particularly interesting examples. I'll then turn to the topic of approximate folding sequences, which, though perhaps not as mathematically interesting, are of considerable practical utility. Along the way, I'll present the axiomatic theory of origami constructions, which not only stipulates what classes of proportions are foldable, but also provides the basis for finding extremely efficient approximate folding sequences by computer solution—a technique that has found application in a number of published origami books of designs.

Preliminaries and Definitions

Origami, like geometric constructions, has many variations. In the most common version, one starts with an unmarked square sheet of paper. Only folding is allowed: no cutting. The goal of origami construction is to precisely locate one or more points on the paper, often around the edges of the sheet, but also possibly in the interior. These points, known as *reference points*, are then used to define the remaining folds that shape the final object. The process of folding the model creates new reference points along the way, which are generated as intersections of creases or points where a crease hits a folded edge. In an ideal origami *folding sequence*—a step-

by-step series of origami instructions—each fold action is precisely defined by aligning combinations of features of the paper, where those features might be points, edges, crease lines, or intersections of same.

Two examples of creating such alignments are shown in Figures 1 and 2. Figure 1 illustrates folding a sheet of paper in half along its diagonal. The fold is defined by bringing one corner to the opposite corner and flattening the paper. When the paper is flattened, a crease is formed that (if the paper was truly square) connects the other two corners.



1. Fold the bottom right corner up to the top left.

2. Unfold.

3. "Fold and unfold" is indicated by a double-headed arrow.

Figure 1. The sequence for folding a square in half diagonally.

As a shorthand notation, the two steps of folding and unfolding are commonly indicated by a single double-headed arrow as in the third step of Figure 1.

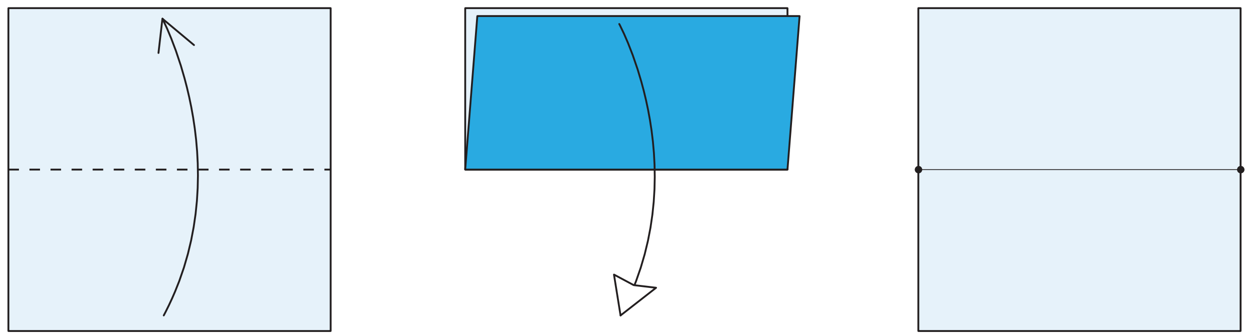
Figure 2 illustrates another way of folding the paper in half ("bookwise"). This fold can be defined in 3 distinct, but equivalent ways:

(1) Fold the bottom left corner up to the top left corner.

(2) Fold the bottom right corner up to the top right corner.

(3) Fold the bottom edge up to be aligned with the top edge.

For a square, these three methods are equivalent. However, if you start with slightly skew paper (a parallelogram rather than a square), you will get slightly different results from the three.



1. Fold the bottom edge up to the top edge.

2. Unfold.

3. The new crease defines two new points.

Figure 2. The sequence for folding a square in half bookwise.

In both cases, if you unfold the paper back to the original square, you will find you have created a new crease on the paper. For the sequence of figure 2, you will also have now defined two new points: the midpoints of the two sides. Each point is precisely defined by the intersection of the crease with a raw edge of the paper.

These two sequences also illustrate the rules we will adopt for origami geometric constructions. The goal of origami geometric constructions is to define one or more points or lines within a square that have a geometric specification (e.g., lines that bisect or trisect angles) or that have a quantitative definition (e.g., a point 1/3 of the way along an edge). We assume the following rules:

(1) All lines are defined by either the edge of the square or a crease on the paper.

(2) All points are defined by the intersection of two lines.

(3) All folds must be uniquely defined by aligning combinations of points and lines.

(4) A crease is formed by making a single fold, flattening the result, and (optionally) unfolding.

Rule (4), in particular, is fairly restrictive; it says that folds must be made *one at a time.* By contrast, all but the simplest origami figures include steps in which multiple folds occur simultaneously. Later in this article, I will discuss what happens when we relax this constraint.

Binary Divisions

One of the most common origami constructions that turns up in practical folding is the problem of dividing one or both sides of the square into N equal divisions, where N is some integer. Figure 2 illustrated the simplest case—dividing the edge of a square into two parts—and its solution. Of course, this method is not restricted to a square; it works equally well on any line segment in a square. Thus, the two halves of the square may be individually divided into two parts, and so on. By repeatedly dividing the segments in half, it is possible to divide the edge of a square (or rectangle) into 4ths, 8ths, and so forth, as shown in Figure 3.

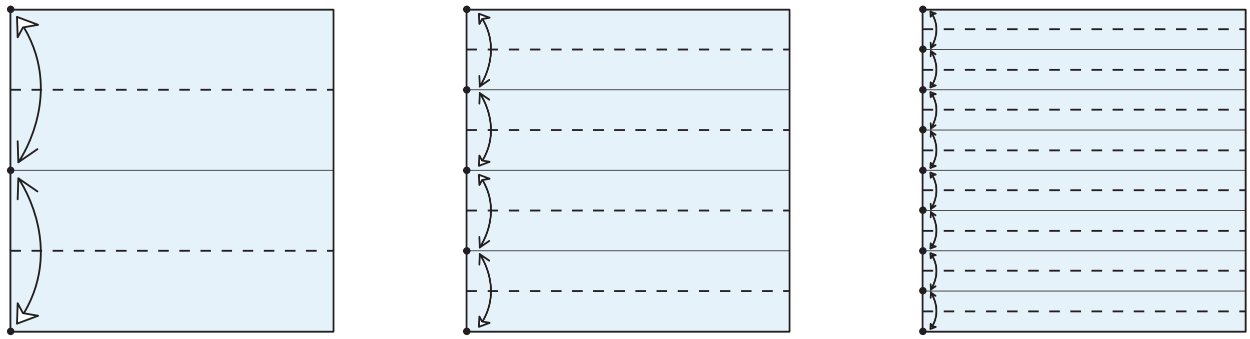


Figure 3. Division of a square into 4ths, 8ths, and 16ths.

This method allows us to divide a square into proportions of 1/2, 1/4, 1/8,…and in general, $1/2^n$ for integer n. Each division is $1/2^n$ of the side of the square. By scaling all numbers to the size of the square, we can say we have constructed the fraction $1/2^n$, where the fraction is given in terms of the side of the square.

It is also possible to construct a fraction of the form $m/2^n$ for any integer $m < 2^n$. (In all the discussion that follows, we will consider only fractions between 0 and 1.) The most direct method is to subdivide the edge of the square completely into 2^nths, then count up m divisions from the bottom. This method clearly requires $2^n - 1$ creases, and is not very efficient, because completely subdividing the square results in the creation of many unnecessary creases. There is an elegant method for constructing any fraction of this type that uses the minimal number of folds. A rational fraction whose denominator is a perfect power of two is called a *binary fraction*; the folding method is called the *binary folding algorithm.*

Binary Folding Algorithm

The binary folding algorithm was described by Brunton [1] and expanded upon by Lang [2]. It produces an efficient folding sequence to construct any proportion that is a binary fraction and is based on binary notation. In binary notation, there are only two digits, 1 and 0; all numbers are written as strings of ones and zeros. Any number can be written in binary notation as a string of ones and zeros. For example, the numbers 1 through 10 can be written in binary as shown in Table 1.

Decimal	Binary
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010

Table 1. Binary equivalents for decimal numbers 1–10.

Any binary fraction of the form $m/2^n$ can be folded in exactly n creases, and the required folding sequence is encoded in the binary expression of the fraction.

Binary notation for fractions is best understood in analogy with ordinary decimal notation. In decimal notation, each digit to the left of the decimal point is understood to multiply a power of 10; for example,

$$1043 = 1\times 10^3 + 0\times 10^2 + 4\times 10^1 + 3\times 10^0 = 1000 + 0 + 40 + 3. \tag{1}$$

The same thing happens in binary notation, except you use powers of 2 rather than powers of 10 and there are only two possible digits: 1 and 0. Therefore, the binary number 1011 is

$$1011 = 1\times 2^3 + 0\times 2^2 + 1\times 2^1 + 1\times 2^0 = 8 + 0 + 2 + 1 = \text{eleven}. \tag{2}$$

By this means, any integer may be written in binary notation with a unique combination of ones and zeros.

While it is less commonly done, it is also possible to write fractional quantities in a binary notation that is analogous to our decimal notation, in which fractional quantities appear as digits to the right of the decimal point (although perhaps it should be called a "binary point" rather than a "decimal point"). For example, just as the decimal 0.753 means

$$.753 = 7\times 10^{-1} + 5\times 10^{-2} + 3\times 10^{-3} = \frac{753}{1000}, \tag{3}$$

the binary fraction .111 may be interpreted as

$$.111 = 1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3} = \frac{7}{8}. \tag{4}$$

Other examples: the fraction 1/2 is given by .1 in binary; the fraction 1/4 is .01 in binary, while 3/4 is .11. The fraction 5/8 is .101, and 23/32, written in binary, is .10111. Any fraction whose

denominator is a perfect power of two has a binary representation with a finite number of digits to the right of the decimal point.

You can construct the binary fraction for any number by following this algorithm:

> (1) Write down a decimal point.
>
> (2) Multiply the fraction by 2.
>
> (3) Subtract off the integer part (either 1 or 0) and write it down to the right of the last thing you wrote.
>
> (4) Repeat steps (2) and (3) as many times as necessary, each time adding digits to the right, until you get a remainder of 0.

Equivalently, the fraction $m/2^n$ is written as a decimal point plus the binary expansion of the integer m, padded with enough zeros to the immediate right of the decimal to get a total of n digits.

What about fractions whose denominator is not a perfect power of 2 (which includes most numbers)? If you write a number such as 1/3 in binary using the algorithm described above, you will never get a remainder of zero. Instead, it forms an infinite string of digits; for example, 1/3=0.010101… If the number is a rational number—the ratio of two integers—then the fraction will eventually start to repeat itself.

The binary expression for a fraction gives a precise description of the folding sequence needed to make a mark at a given distance up the side of the paper. First, here's the folding algorithm:

> To mark off a distance equal to a binary fraction by folding, write down its binary form.
>
> Then, beginning from the *right* side of the fraction (the least significant digit): for the first digit (which is always a 1 because you drop any trailing zeros) fold the top down to the bottom and unfold.
>
> For each remaining digit, if it is a 1, fold the top of the paper to the previous crease, pinch, and unfold; if it is a 0, fold the bottom of the paper to the previous crease, pinch, and unfold.

By comparing this algorithm with the expansion formula for a binary fraction, you can see how the folding algorithm works. Let's take the number 0.11001 (25/32) as an example. The conventional way of expanding this is to expand the number in powers of 2, as shown in equation (5).

$$0.11001 = 1\times 2^{-1} + 1\times 2^{-2} + 0\times 2^{-3} + 0\times 2^{-4} + 1\times 2^{-5} \tag{5}$$

Another way of writing this binary expansion is to expand it as a nested series, as in equation (6).

$$0.11001 = \frac{1}{2}\times(1+\frac{1}{2}\times(1+\frac{1}{2}\times(0+\frac{1}{2}\times(0+\frac{1}{2}\times(1))))) \tag{6}$$

To evaluate this form, you start at the innermost number in the expression (the terminal "1") and work your way back to the left, slowly working your way out of the nested parentheses. If we write the fraction this way, it becomes a series of nested operations where each operation is either:

(a) Add 0 and multiply by 1/2, or

(b) Add 1 and multiply by 1/2.

Now let's look at the origami folding sequence in the recipe above. If we have a square with a crease mark located a distance r from the bottom and fold the bottom of the square up and unfold, the new crease is made a distance $(1/2)r$ from the bottom. If instead, we fold the top of the square down to the mark and unfold, the new crease is made a distance $(1/2)(1+r)$ from the bottom. Thus, folding the bottom up or top down is equivalent to performing operations (a) or (b), respectively.

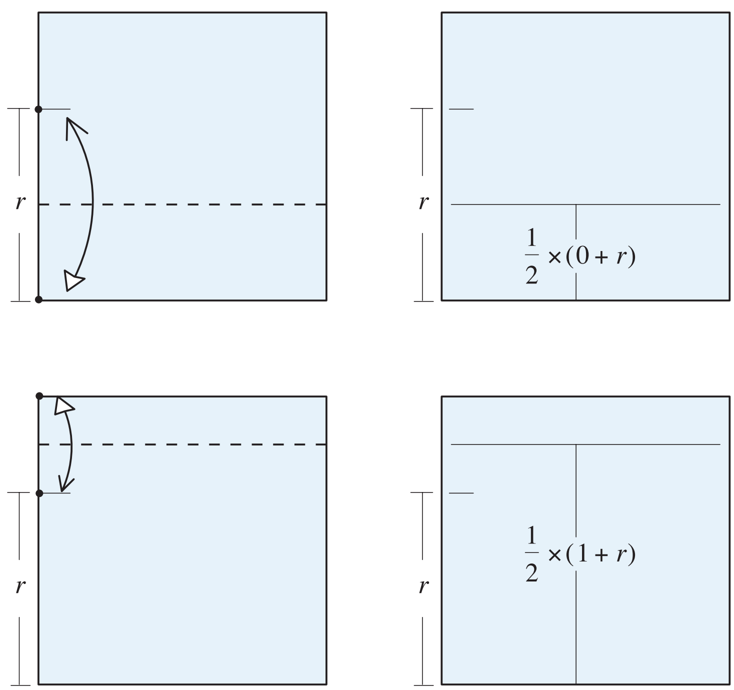


Figure 4. (Top) Folding the bottom edge up to a crease r gives a new crease $(r/2)$ from the bottom. (Bottom) Folding the top edge down to a crease r gives a new crease $((1+r)/2)$ from the bottom.

Since any binary fraction can be written as a nested sequence of the two operations (a) and (b) and the two folding steps shown in figure 1 implement these two operations, it follows that any proportion can be folded from its binary expansion.

The difference in efficiency between folding all divisions and counting upward, versus the binary method, is substantial. For a fraction $m/2^n$, the former method requires $2^n - 1$ folds; the latter, only n.

Binary Approximations

Only fractions whose denominator is a perfect power of 2 possess a binary expansion with a finite number of digits. For most fractions, the binary expansion of the fraction is infinite. But if we truncate the binary expansion at some point, we get a binary fraction that provides a close approximation of the number. This works in any number base. For example, in decimal notation, 1/3=0.3333… (also an infinite decimal). If we truncate at one digit (0.3), we get the fraction 3/10, which is only roughly equal to 1/3. If we take two digits (0.33), we get 33/100, which is very close to 1/3; and if we take 3 digits (0.333), we get 333/1000, which is very close indeed.

The same thing happens in binary notation. If we truncate the binary expansion of 1/3 at 2 digits, we get 0.01=1/4 — a rather crude approximation of 1/3. But 0.0101 is 5/16, which is closer to 1/3, and 0.010101 is 21/64, which differs from 1/3 by less than 1%. Thus, any number can be approximated by a binary fraction to arbitrary accuracy, which leads to an easy way to find an approximation of any proportion by folding: Construct the binary expansion of the fraction; truncate the expansion at a desired level of accuracy; then use the binary algorithm to construct a folding sequence.

Fractions that are the ratio of two integers where the denominator is not a power of 2 have binary expansions that eventually repeat. This property allows an iterative folding sequence that successively approximates the desired proportion. The repeating part defines the folding sequence that is to be repeated

For example, the binary expansion of 1/3 is $.\overline{01}$, where the overbar indicates repetition (i.e., $.\overline{01} = .010101\ldots$). The repeating part, 01, defines the sequence (remember, we start at the right), "Fold the top down to the previous mark and unfold; fold the bottom up to the previous mark and unfold." Repeating this procedure over and over will produce a series of pairs of crease marks that fairly rapidly converges on 1/3 and 2/3, as illustrated in Figure 5.

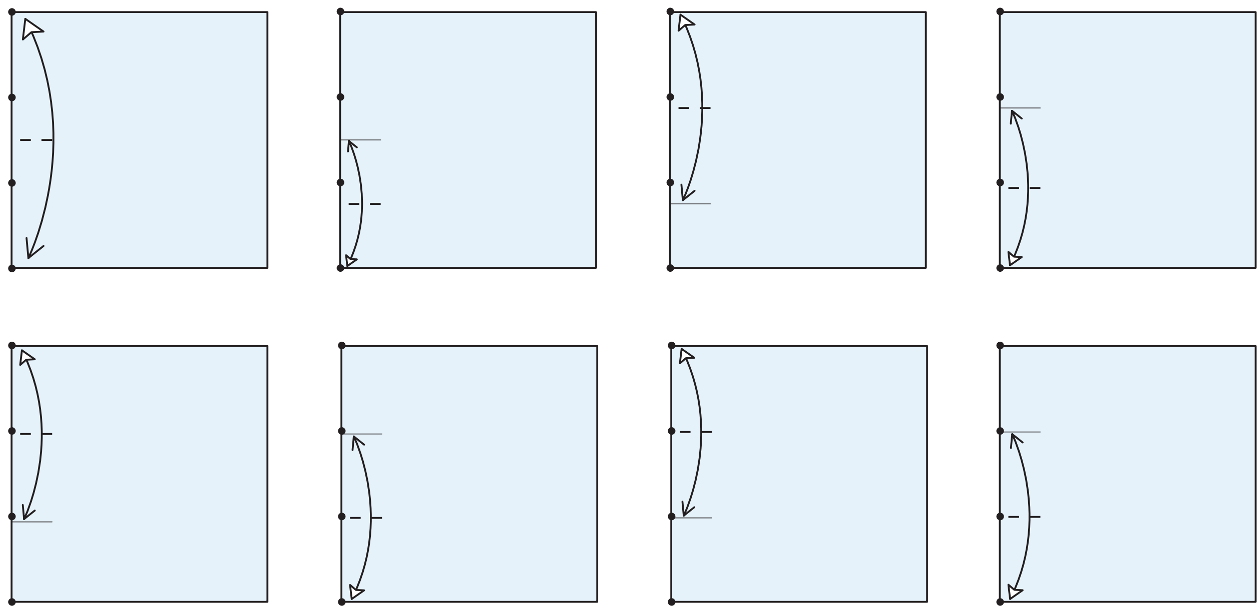


Figure 5. Iterative folding sequence to find 1/3.

A similar iterative technique exists for finding 1/5, whose binary expansion is $.\overline{0011}$. Its iterative sequence, too, can be read off from its binary expansion: fold the top down twice, then the bottom up twice; repeat as needed. Since all non-binary rational fractions eventually repeat, there are iterative procedures for them all.

One can also consider the converse; suppose we choose a procedure, like "fold the top down twice, then the bottom up three times; repeat." What fraction does this converge to? Such a procedure would have a binary expansion of $.\overline{11000}$. There is a well-known procedure for converting a repeating expansion into a rational fraction. You writing the repeating part in the numerator, and fill the denominator with the same number of digits *d*, where *d* is one less than the base of the number system. In our example, *d*=1, and thus

$$.\overline{11000} = \left.\frac{11000}{11111}\right|_{binary} = \left.\frac{24}{31}\right|_{decimal}. \qquad (7)$$

The iterative procedure for 1/3 shown in Figure 5 converges on two creases, at 1/3 and 2/3 of the way along the edge. That's because the iterative procedure defined by 01 corresponds to two repeating fractions: $.\overline{01}$ and $.\overline{10}$, whose repeating parts are cyclic permutations of one another. By the same token, it should be clear that any repeating folding sequence will converge to the set of creases defined by all cyclic permutations of the repeating part. Thus, for example, 011 (down, down, up) will converge to creases at

$$\frac{001}{111} = \frac{1}{7},\ \frac{010}{111} = \frac{2}{7},\ \text{and}\ \frac{100}{111} = \frac{4}{7}. \qquad (8)$$

Since any number, rational or not, can be approximated by a binary expansion, this technique gives a way of folding any proportion to arbitrary accuracy.

The power of the binary approximation algorithm is that it attains fairly good accuracy with a relatively small number of folds. One can easily compute the number of folds necessary to attain a given level of accuracy. If you want to fold a fraction r to an accuracy ε, the number of creases required by a binary approximation is less than or equal to

$$\left\lceil \left(\log_2 \frac{1}{\varepsilon}\right) - 1 \right\rceil, \tag{9}$$

where $\lceil \ldots \rceil$ is the ceiling function (round upward to the nearest integer).

The number of creases needed to fold a given proportion is an important practical measure of a folding sequence, called the *rank* of the sequence. A low rank takes less time and in general, leaves fewer unnecessary creases on the paper. For a finite binary fraction *m*/*p* (reduced to lowest terms), it is clear that the rank of the binary fold method, denoted by *bin*(*m*/*p*), is given by

$$rank\big(bin(m/p)\big) = \log_2 p. \tag{10}$$

From a purely mathematical standpoint, constructions that are mathematically exact are most interesting, but from a practical standpoint, approximate constructions with low rank are more useful. To get one-part-in-a-thousand accuracy (more accurate that is usually required in real-world origami), equation (9) shows that we would need no more than 9 creases to approximate the desired proportion. In practice, the number of creases can be less than the theoretical maximum. Some proportions will just happen to have binary expansions that are accurate with fewer than 9 digits.

Another nice property of the binary algorithm is that you can make most of the creases with small pinch marks along the edge of the paper; it doesn’t clutter up the main square with a lot of extraneous creases.

There is another use for the binary algorithm; it is a key element in several exact distance-finding algorithms. While the binary algorithm is exact only for fractions whose denominator is a perfect power of two, there are several other algorithms that can fold any rational fraction exactly. These algorithms are described in subsequent sections.

Rational Fractions

In the style of folding known as box-pleating, typified by the works of Hulme and Elias, among others, the paper is initially creased into a grid of equal-sized squares. A model might begin by dividing the paper into twelfths, sixteenths, or less commonly, ninths, fifteenths, or even such oddities as 78^{th}s [3]. The frequency of the need to divide a square into a set number of equal divisions leads to a mathematical construction problem: how to divide a square into *b* equal parts. More generally, we can ask the question, how can we construct by folding alone a segment of length *a*/*b* times the side of the square, where *a* and *b* are both integers and *b* is *not* a power of 2. The binary algorithm lets us find any fraction of the form *m*/*p*, where *p* is a power of 2. Is it possible to start with one or more binary fractions and construct proportions equal to non-binary fractions? There are several different ways of doing this.

Crossing Diagonals

The construction for one of the most versatile origami constructions for an arbitrary fraction a/b is shown in Figure 6. It uses two creases: one of them is the diagonal of the square; the other is a crease that connects two points on opposite sides.

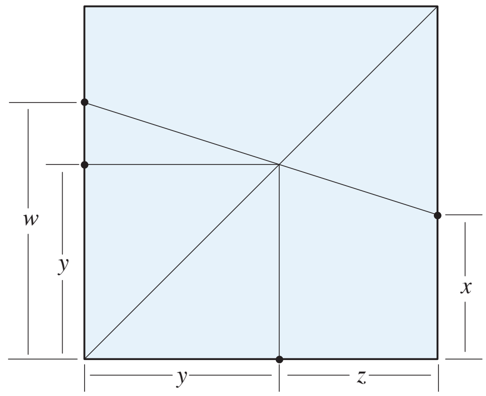


Figure 6. Construction for finding a rational number as the fraction of the side of a square.

We start with a unit square in which we have creased the diagonal that runs from lower left to upper right. We then construct two marks at distances w and x, respectively, along each of the two sides, and connect them with a crease. The intersection of the two creases defines a new point, whose projection onto any edge defines a new distance y. Solving for y and its complement $z=1-y$, gives

$$y=\frac{w}{1+w-x},\ z=\frac{1-x}{1+w-x}. \tag{11}$$

The idea behind the crossing-diagonals construction (and many others) is that one picks the two initial proportions w and x to be relatively easy to construct, i.e., binary fractions, in order to construct the fraction y (or z), which is a non-binary fraction (which we will denote by a/b). Thus, we take w and x to be the binary fractions

$$w\equiv\frac{m}{p},\ \ x\equiv\frac{n}{p}, \tag{12}$$

where m and n are integers smaller than p and p is a power of 2. Then

$$y=\frac{m}{p+m-n},\ z=\frac{p-n}{p+m-n}. \tag{13}$$

Setting $y=a/b$ gives rise to the following sequence.

Define p to be the next power of 2 equal to or larger than both a and $b–a$.

Define $m=a$, $n=(p+a–b)$.

Construct the points $w=m/p$, $x=n/p$ along the left and right edges using the binary method. Connect them with a crease.

Construct the diagonal.

The intersection of the two creases defines the fraction a/b as its height above the bottom of the square (or equivalently its distance from the left edge).

Let's look at a few examples. The most common odd division of a square is to divide it into thirds. If we take a/b=1/3, then p=4, m=1, n=2, which gives rise to the folding sequence shown in Figure 7.

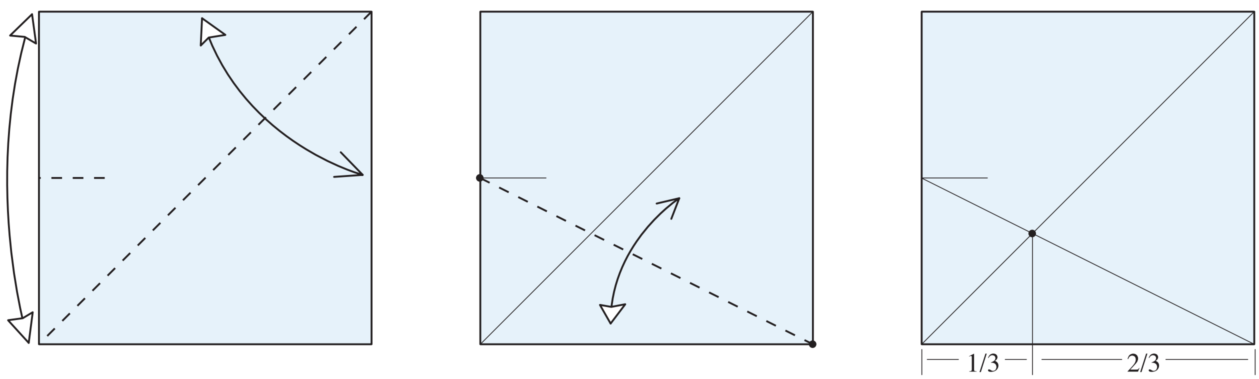


Figure 7. An exact folding sequence for dividing a square into thirds.

The sequence for dividing into thirds shown in figure 7 is quite well-known in origami. It is just one example of a general origami construction, known as the *crossing diagonals method* [2], which can be applied to any non-binary rational. Table 2 tabulates the values of w and x, as well as the rank, for the reduced non-binary fractions with denominators up to 10. (Note that for a fraction $y=a/b$, the distance marked z in Figure 6 gives the fraction $(b–a)/b$, so we only need to consider fractions smaller than 1/2.)

y=a/b	*z=1–y*	*w*	*x*	*rank*
1/3	2/3	1/2	0	3
1/5	4/5	1/4	0	4
1/6	5/6	1/8	3/8	8
1/7	6/7	1/8	1/4	7
2/7	5/7	1/4	3/8	7
3/7	4/7	3/4	0	4
1/9	8/9	1/8	0	5
2/9	7/9	1/4	1/8	7
4/9	5/9	1/2	3/8	6
1/10	9/10	1/16	7/16	10
3/10	7/10	3/8	1/8	8

Table 2. Reduced non-binary fractions and the binary fractions that give rise to their construction.

There are many possible variations on this basic idea for finding rational number proportions. They are all based on the idea of crossing two diagonal creases that have different slopes. (The same concept can also be applied to find many irrational numbers, notably bilinear combinations of integers and $\sqrt{2}$, as we will see later.) Here's another version of crossing-diagonals. Instead of taking one crease always to be the diagonal of the square and the other connecting two points on opposite sides, one could instead cross two diagonals, both of which begin from the bottom corners of the square, as illustrated in Figure 8.

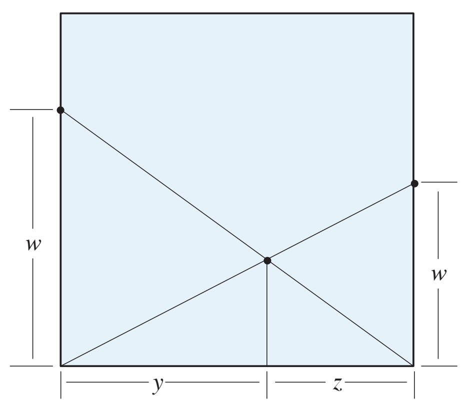


Figure 8. An alternative crossing diagonals construction for finding proportions.

For this construction, we find that the bottom edge is divided into the fractions

$$y = \frac{w}{w + x}, \; z = \frac{x}{w + x}. \tag{14}$$

Again, choosing our proportions w and x to be binary fractions,

$$w \equiv \frac{m}{p}, x \equiv \frac{n}{p}, \tag{15}$$

we find that

$$y = \frac{m}{m+n},\ z = \frac{n}{m+n}. \tag{16}$$

This gives rise to the folding sequence below for a fraction *a*/*b*.

Define *p* to be the smallest power of 2 larger than both *a* and *b*-*a*.

Define *m*=*a*, *n*=*b*–*a*.

Construct the points *w*=*m*/*p*, *x*=*n*/*p* along the left and right edges using the binary method.

Connect points *w* and *x* with the bottom opposite corners with creases.

The intersection of the two creases defines the fraction *a*/*b* as its height above the bottom of the square (or equivalently its distance from the left edge).

Table 3 gives the construction fractions and ranks for the same fractions as in Table 2. It turns out that for a given fraction, the two crossing diagonals methods have the same rank.

y=a/b	*z=1–y*	*w*	*x*	*rank*
1/3	2/3	1/2	1	3
1/5	4/5	1/4	1	4
1/6	5/6	1/8	5/8	8
1/7	6/7	1/8	3/4	7
2/7	5/7	1/4	5/8	7
3/7	4/7	3/4	1	4
1/9	8/9	1/8	1	5
2/9	7/7	1/4	7/8	7
4/9	5/9	1/2	5/8	6
1/10	9/10	1/16	9/16	10
3/10	7/10	3/8	7/8	8

Table 3. Construction fractions and rank for the second crossing diagonals folding sequence.

Fujimoto's Construction

An alternative technique for folding rational fractions was devised by the Japanese mathematician Shuzo Fujimoto [4] and was independently rediscovered by the Boston geometer Jeannine Mosely [5]. Fujimoto's algorithm relies on an elegant construction for taking reciprocals of folded proportions, based on the construction shown in Figure 9.

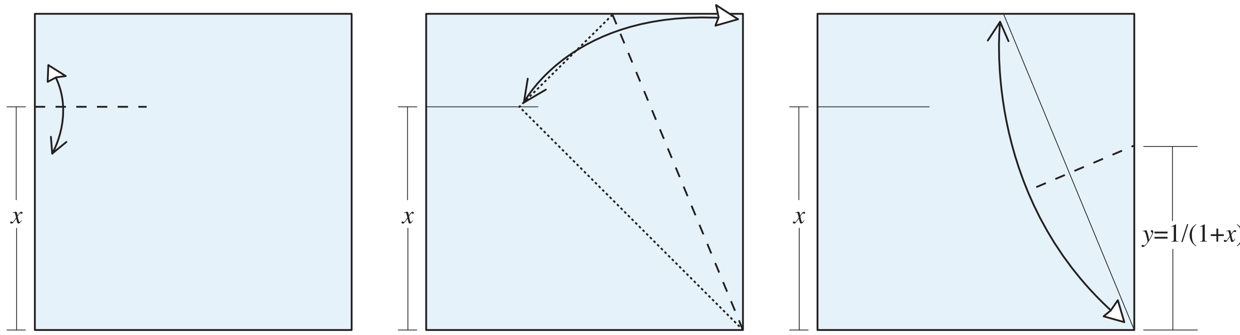


Figure 9. Schematic of Fujimoto's construction of a reciprocal.

Beginning from a proportion x defined by a crease along one side of a square, this two-fold sequence produces the reciprocal of $(1+x)$. So, for example, if you want to find the reciprocal of a number y, if you start with the proportion $(y-1)$ marked off along the left side, Fujimoto's construction will produce the number $1/(1+y-1)=1/y$.

To construct a fraction a/b, we define x to be a binary fraction

$$x \equiv \frac{m}{p}. \tag{17}$$

Using the Fujimoto construction, the distance y is

$$y = \frac{p}{m+p}. \tag{18}$$

We take p to be the largest power of 2 smaller than the denominator b, and $m=b-p$. Then

$$y = \frac{p}{b}, \tag{19}$$

which gives the desired denominator b. Since p is a power of 2, we can use the binary algorithm to reduce this fraction by the factor (a/p), giving the final proportion:

$$z = \frac{a}{p} y = \frac{a}{p} \times \frac{p}{b} = \frac{a}{b}. \tag{20}$$

The complete algorithm is summarized below.

Define p as the largest power of 2 smaller than b.

Define $x=(b-p)/p$.

Construct x using the binary algorithm, extending the final horizontal crease as shown in Figure 9.

Apply Fujimoto's construction. This will give the fraction (p/b) along the right side of the paper, defined by the mark along the right.

Reduce this distance by the fraction a/p, again, using the binary algorithm.

I summarize the construction fractions and rank for the irreducible non-binary fractions in Table 4.

y	*1–y*	*x*	*a/p*	*rank*
1/3	2/3	1/2	1/2	4
1/5	4/5	1/4	1/4	6
1/6	5/6	1/2	1/4	5
1/7	6/7	3/4	1/4	6
2/7	5/7	3/4	1/2	5
3/7	4/7	3/4	3/4	6
1/9	8/9	1/8	1/8	8
2/9	7/9	1/8	1/4	7
4/9	5/9	1/8	1/2	6
1/10	9/10	1/4	1/8	7
3/10	7/10	1/4	1/4	6

Table 4. Construction fractions and rank for Fujimoto's algorithm.

Although both crossing-diagonals and Fujimoto's algorithms provide exact folding techniques for any rational fraction, the folding sequence may be imprecise in practice, for example, requiring one to fold a long, skinny triangular flap (which is difficult to do neatly). The various construction methods are sometimes complementary; when one algorithm is lengthy, the other may be short, and when one is imprecise, the other is not. For comparison, a division into equal fifths is shown in Figures 10 and 11 for two methods.

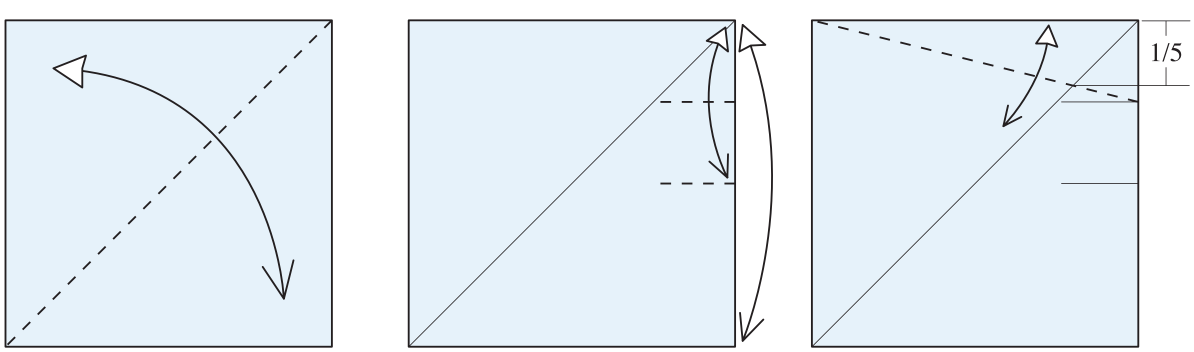


Figure 10. Crossing diagonals algorithm for division into fifths.

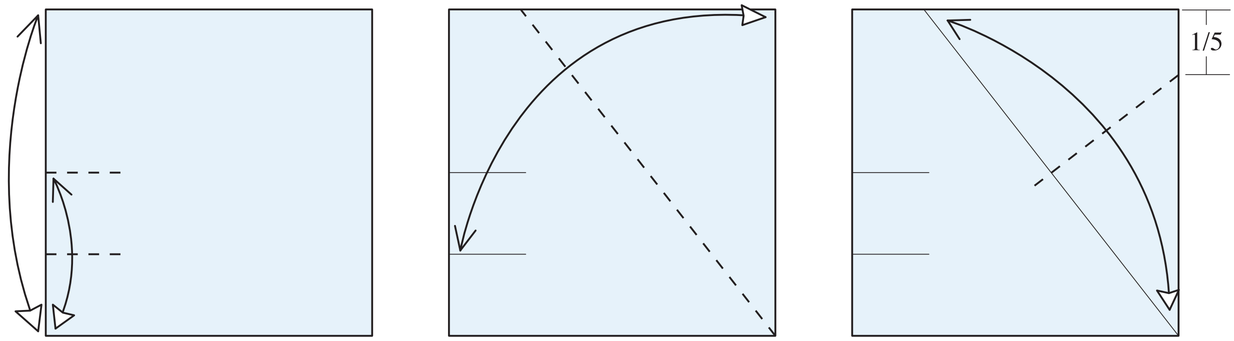


Figure 11. Fujimoto's algorithm for division into fifths.

One drawback of the crossing-diagonals and Fujimoto algorithms is that they leave extra creases running across the middle of the paper. Wouldn't it be nice, though, if there were a construction that could produce any possible fraction and that was constructed only with pinch marks around the edge and put no creases in the interior of the paper? There is such a construction, and it is the subject of the next section.

Noma's Method

If you start with the requirement that the only allowed creases are pinch marks around the edges, you quickly find that there are only a few possible types of fold that create new marks on the edges. The two simplest are:

(1) You can bring one mark on an edge to another mark on the same edge. This is what we do when we use the binary division algorithm; and we know already that this will only provide fractions whose denominators are powers of 2.

(2) You can bring one mark on an edge to a different mark on a different edge.

There are others (which we will encounter later), but there is substantial unrealized potential in just these two operations. Consider the case where we bring together two marks on adjacent edges and make new marks where the resulting crease hits the edges, as shown in Figure 12. The relevance of this operation to origami constructions was discovered by Masamichi Noma [6], and so we will call it Noma's construction.

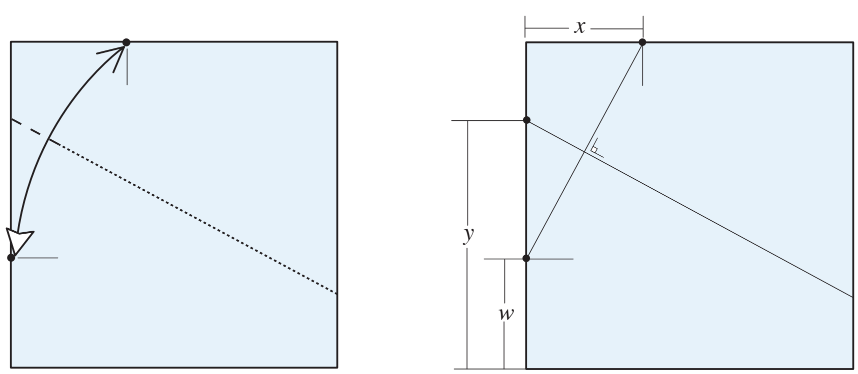


Figure 12. Schematic of Noma's construction.

By working out the various dimensions (some of which are shown in Figure 12), one can show that if one takes

$$w = x = \frac{b}{2p}, \tag{21}$$

then the point y is a distance

$$y = \frac{p}{b} \tag{22}$$

above the bottom of the square. This leads to the following algorithm.

Define p as the largest power of 2 smaller than b.

Construct the fractions $w=b/2p$, $x=b/2p$ along the left and top edge, respectively.

Bring point w to point x, making a crease along the left edge at height $y=p/b$.

Construct the fraction a/p relative to this segment.

The result is the desired fraction a/b.

The full algorithm is illustrated in the abbreviated folding sequence shown in Figure 13.

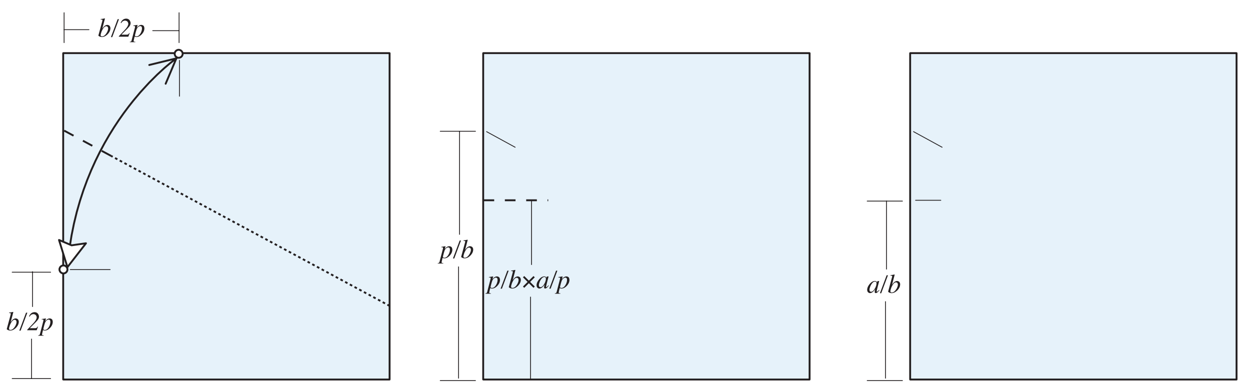


Figure 13. The complete Noma algorithm for any rational fraction.

The required fractions and ranks for the rationals with denominators up to 10 are given in Table 5.

y	*1–y*	*b/2p*	*a/p*	*rank*
1/3	2/3	3/4	1/2	6
1/5	4/5	5/8	1/4	9
1/6	5/6	3/4	1/4	7
1/7	6/7	7/8	1/4	9
2/7	5/7	7/8	1/2	8
3/7	4/7	7/8	3/4	9
1/9	8/9	9/16	1/8	12
2/9	7/7	9/16	1/4	11
4/9	5/9	9/16	1/2	10
1/10	9/10	5/8	1/8	10
3/10	7/10	5/8	3/8	10

Table 5. Fractions, construction fractions, and rank for Noma's algorithm.

There is a tradeoff here; we need to apply the binary algorithm three times (first to the two different edges, then again to divide down the Noma division), so that the rank of Noma's method is generally higher than the rank of the other methods.

Haga's Construction

Yet another construction was discovered by Kazuo Haga [7–9], which requires only a single diagonal crease and can also produce all rational fractions. The construction is generally known as "Haga's theorem." A variation of Haga's theorem, discovered by Husimi, also provides a division into fifths, which should be compared with the two previous examples of division into fifths. It is shown in Figure 14.

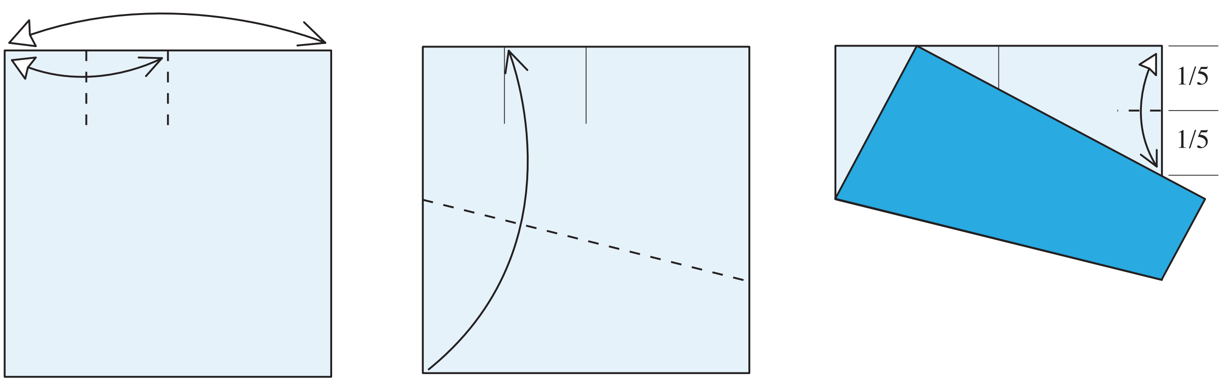


Figure 14. A division into fifths based on the Haga theorem.

Like the other two algorithms, there are numerous variations of Haga's construction for finding other proportions that are rational fractions. The general form of the Haga construction is shown in Figure 15. There are two variations; the desired reference point can be the crossing of the two raw edges, in which case the mark is formed by folding along one of the two edges, as in the middle image of Figure 15. In the second, one folds the upper corner to the intersection.

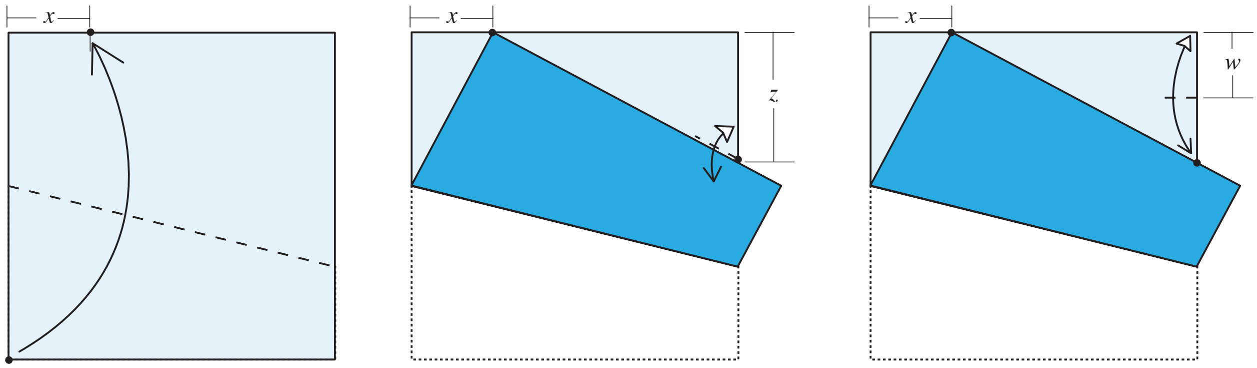


Figure 15. Schematic of the general Haga construction.

Haga's construction differs from the others in that the paper is not unfolded between all folds. However, it permits some particularly efficient rational constructions. If we make the first fold at a distance x along the top edge, then the two constructed distances in Figure 15 are

$$z = \frac{2x}{1+x}, \quad w = \frac{x}{1+x}. \tag{23}$$

This leads to the following construction for a fraction a/b.

> Define p to be the largest power of 2 smaller than b.
>
> Define $m=p–b$.
>
> Construct the point $x=m/p$ along the top edge using the binary method.
>
> Fold the bottom left corner up to the top edge.
>
> Fold the top right corner down to the crossing of the two raw edges and unfold, defining the distance $y=p/b$.
>
> Reduce the segment y by the fraction a/p using the binary method. The result is the desired fraction a/b.

These dimensions are illustrated in Figure 16.

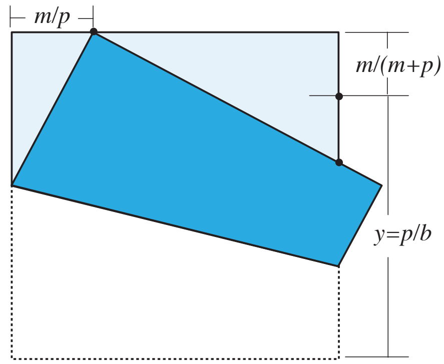


Figure 16. Relevant dimensions for the construction of the fraction *a/b* using Haga's construction.

With the Haga construction, the diagonal crease doesn't need to be made sharp anywhere along its length; the edge of the fold only needs to be held down while folding down the upper right corner that defines the distance *w*. Table 6 gives the relevant fractions for constructions using the Haga construction and their rank.

y=a/b	*1–y*	*m/p*	*a/p*	*rank*
1/3	2/3	1/2	1/2	4
1/5	4/5	1/4	1/4	6
1/6	5/6	1/2	1/4	5
1/7	6/7	3/4	1/4	6
2/7	5/7	3/4	1/2	5
3/7	4/7	3/4	3/4	6
1/9	8/9	1/8	1/8	8
2/9	7/7	1/8	1/4	7
4/9	5/9	1/8	1/2	6
1/10	9/10	1/4	1/8	7
3/10	7/10	1/4	3/8	7

Table 6. Irreducible fractions, their construction fractions, and rank for Noma's method.

These solutions are, in general, simpler than the Noma construction, and if the diagonal crease is not pressed flat, can also be made without marking the interior of the paper.

Irrational Proportions

Continued Fractions

While many geometric constructions are possible with origami and many proportions can be folded exactly, there are other proportions for which an exact folding sequence is either impossible with origami (like 1/π) or even if it is possible, it may leave the paper covered with so many creases as to be wholly impractical for any real folding. To the practicing origami artist, the question is not "how can I fold this proportion exactly?" but "how can I fold this proportion to necessary accuracy in as few creases as possible?" Ideally, one would find a mathematically

exact method for folding the distance, but mathematical exactitude isn't always necessary. In real-world folding, distance errors of less than 0.5% of the side of the square are rarely discernible. Consequently, one doesn't have to find an exact method for folding a proportion: it merely suffices to find a method of folding a close approximation of the proportion.

Here is a simple example; suppose we wished to construct a 60° angle inside one corner of a square, creating a 30–60-90 right triangle on one side. One way of doing this would be to locate the point where the crease intersects the side of the square, as shown in figure 17. Since the sides of such a triangle are in the proportions 1:√3:2, expressed as a fraction of the side of the square, the distance from the corner to the crease along the bottom is the quantity 1/√3=0.577…. One way of constructing the angle is to find the point along the bottom where the line hits it, that is, to find the distance 1/√3. This distance is neither a binary fraction nor a rational fraction, so we don't currently know an exact solution. How can we find a rational fraction approximation to this number that is accurate to better than a specified tolerance?

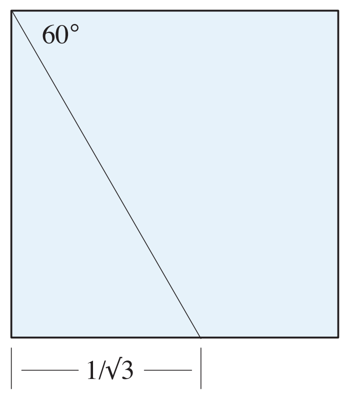


Figure 17. One way of constructing a 60° angle is to mark off a distance 1/√3 along one side of the square.

(Note: there happen to be several elegant and exact constructions for finding a 60° angle, but we'll overlook them for the moment for purposes of illustration.)

The most direct way to fold a proportion is the brute-force one; write the number as a decimal, for example, 1/√3=0.57735…. Truncate it at three digits and write the decimal as a fraction;

$$\frac{1}{\sqrt{3}} = 0.57735\ldots \approx 0.577 = \frac{577}{1000}. \tag{24}$$

Divide the paper into one-thousandths, and count off five hundred and seventy-seven divisions.

While this is clearly brute-force and inelegant, the binary algorithm described in the first section works in approximately the same fashion. If we write this fraction in binary, we get

$$\frac{1}{\sqrt{3}} = 0.1001001111\ldots \approx \frac{591}{1024}, \tag{25}$$

and we could apply the binary algorithm (ten consecutive pinch marks) to find the desired proportion. But ten pinch marks is a lot of folding. Wouldn't it be nice if we could find a relatively small fraction that still provides a close approximation to the number in question? Often there is, but how to find it?

The answer lies within a mathematical object called a "continued fraction," which arises in number theory and analysis [10]. A continued fraction is a way of representing a number as a fraction within a fraction within a fraction…and so forth. The general form of a continued fraction is

$$r = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \ldots}}}, \tag{26}$$

where r is the number in question and b_0, b_1, and b_2 are (usually) integers. Some continued fractions have a finite number of terms; in others, the nested fractions go on forever. Any number may be written as a continued fraction; in fact, there are infinitely many continued fractions that can represent the same number. However, if we require that the numbers $\{b_n\}$ be positive integers, then the continued fraction representation for a given number is unique — meaning that there's only one sequence of digits you can plug into the fraction to obtain the number. For example, the fraction 3/16 is given by the continued fraction

$$\frac{3}{16} = 0 + \cfrac{1}{5 + \cfrac{1}{3}} \tag{27}$$

which is quite simple. On the other hand, the fraction $1/\sqrt{3}$ is given by the infinite continued fraction

$$\frac{1}{\sqrt{3}} = 0.577\ldots = 0 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \ldots}}}}}} \tag{28}$$

where the ellipsis indicates that the hierarchy of fractions keeps going — forever. If the number r is a rational number — that is, it can be expressed as the ratio between two integers, like 3/16 — there is a finite number of terms in the fraction. If the number is irrational (for example, $1/\sqrt{3}$), the sequence never stops. If the number is the sum of a rational number and the square root of a rational number, it eventually repeats (notice the repeating pattern of 1s and 2s in the fraction above) but for most irrational numbers, the sequence marches on its merry way, *ad infinitum.*

The utility of a continued fraction is this: even if the continued fraction goes on forever, if you chop off the bottom of the infinite fraction, you get a finite fraction that is a close approximation of the original number. The more terms you take, the better is your rational approximation.

With a pocket calculator, it is very simple to determine the first few terms of the continued fraction sequence for any number. Let us take the mathematical constant π=3.1415926535… as an example. Here's how you make a continued fraction:

(1) Subtract the integer part and write it down (*e.g.*, subtract 3, leaving 0.14159…).

(2) Take the reciprocal of the remainder (*e.g.*, 1/0.14159…=7.06251…).

(3) Repeat steps (1) and (2) on the remainder until the remainder is zero or you get tired (or you exceed the resolution of your calculator).

The sequence of integers that you wrote comprises the continued fraction sequence. For the number π, you will find that its sequence is $\pi = \{3;7,15,1,293,10,3,\ldots\}$, which means that

$$\pi = 3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{293 + \cfrac{1}{10 + \ldots}}}}}. \tag{28}$$

If you chop off the bottom of the fraction, you get a rational fraction that is an approximation to the irrational number π. The accuracy of the approximation depends on where you chop the infinite fraction. The first four fractions for π are, for example,

$$3 = 3.00, \tag{29}$$

$$3 + \frac{1}{7} = \frac{22}{7} = 3.1428\ldots, \tag{30}$$

$$3 + \cfrac{1}{7 + \cfrac{1}{15}} = \frac{333}{106} = 3.141509\ldots, \tag{31}$$

$$3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1}}} = \frac{355}{113} = 3.14159292\ldots. \tag{32}$$

$$3 + \cfrac{1}{7 + \cfrac{1}{15 + \cfrac{1}{1 + \cfrac{1}{293}}}} = \frac{104,348}{33,215} = 3.14159265\ldots \tag{33}$$

As you can see from this example, the farther you continue the fraction before chopping it off, the more accurate the rational approximation. The fractions obtained by this procedure are known as *convergents* of the continued fraction. (Recreational mathematicians will recognize 355/113, a famous approximation to π, as the fourth convergent.)

Although you can evaluate the convergents by repeatedly simplifying the complex hierarchical fractional expression, there is a little table that you can construct to quickly evaluate the convergents. Write the continued fraction sequence in the top row of a table as shown in Table 7.

		3	7	15	1	293	…
0	1						
1	0						

Table 7. Convergents for the continued fraction expansion of π.

The first two entries in the next two rows are, respectively, 0, 1 and 1, 0. Then you successively fill in each cell of the next two rows according to this rule:

> The number in any cell is the sum of the number 2 cells to the left and the product of the number at the top of the column with the number to the immediate left.

Using this rule, you fill in the cells from left to right. For example, the cell immediately under the 3 gets filled in with $3\times1+1=3$. The cell below it gets $3\times0+1=1$. The cell immediately under the 7 gets $7\times3+1=22$, and the cell under that gets $7\times1+0=7$. And so forth. For the continued fraction sequence for π, the table fills in as such:

		3	7	15	1	293	…
0	1	3	22	333	355	104,348	…
1	0	1	7	106	113	33,215	…

Table 8. Convergents for the continued fraction expansion of π.

As you can see by comparing this table to the fractions earlier, each convergent is simply the ratio of a number in the middle row and the number below it.

So why go to all this trouble to get a rational approximation; why not just write the number as a truncated decimal? The reason to use continued fractions as rational approximations stems from a unique property of the convergents; each convergent has the smallest possible denominator for a given level of accuracy. Each convergent is the best approximation you can find until the next convergent, where "best" means the smallest possible error. So 22/7 is the best approximation to π with a denominator smaller than 106; 333/106 is the best approximation with a denominator smaller than 113; and 355/113 is the best approximation with a denominator smaller than 33,215, which is anomalously good (which is one reason why this particular fraction is so famous). Continued fraction convergents with small denominators can be very accurate indeed. Even a fraction as simple as 22/7 differs from π by only 0.001.

Even for origami constructions that do not have exact folding sequences, it is possible to come arbitrarily close to the exact proportion using continued fractions. Whatever the number, you need simply to write it as a continued fraction, work out the first 4 or 5 convergents, and pick the smallest convergent that gives an acceptably small error. The problem is thereby simplified; instead of being prepared to find a folding sequence for any number whatsoever, we need only to find a folding sequence for any rational fraction — a ratio of two integers. These can be provided by the folding algorithms already described.

Quadratic Surds

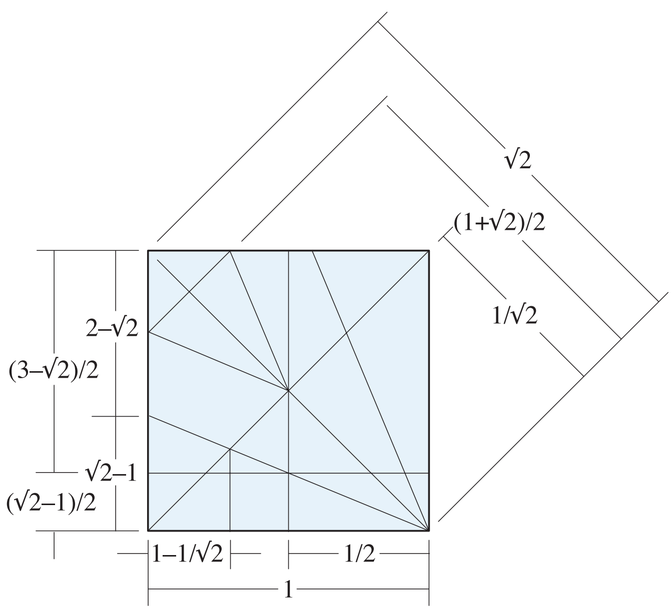
The algorithms I've described thus far apply to rational numbers, ratios of two integers. Sometimes these are required directly, for example, when you must divide the square in ninths; sometimes, we use a rational fraction as an approximation of another proportion. These other proportions may involve square roots, cube roots, trigonometric functions, or may even be numerical values solved for by calculator or computer. All such proportions can be approximated

by converting them to rational numbers and then using an exact folding sequence for the rational proportion.

However, there is another family of irrational proportions that frequently arise within origami for which simple and exact folding solutions often exist: those are proportions of the form

$$\frac{1}{a + b\sqrt{2}} \tag{34}$$

where a and b are integers, which are usually small [2]. Such proportions are called quadratic surds. (To be precise, they are a subset of the quadratic surds; general quadratic surds can have numerators other than 1 and other numbers inside the square root.) These proportions arise often enough within origami that they are worth special mention. Many origami crease patterns make use of symmetries associated with geometric figures whose angles are multiples of 22.5°, which is 1/16th of a unit circle. In such bases, most of the major lines in the crease pattern are proportional to each other by factors that are of the form $a + b\sqrt{2}$. For example, a square with a handful of these angle-bisector creases contains a family of lines forming an ascending series of proportions that are all of this type.



Algebraic	Decimal	Linear
√2	1.414	
(1+√2)/2	1.207	
1	1.000	
(3–√2)/2	0.793	
(1/√2	0.707	
2–√2	0.586	
1/2	0.500	
√2–1	0.414	
1–1/√2	0.293	
(√2–1)/2	0.207	

Figure 18. Bilinear surds that appear in a creased square.

The crease patterns of origami bases that utilize the symmetries of 22.5° geometry are composed of two types of triangles : the 45–45–90 right triangle and the 22.5–67.5–90 right triangles, whose sides have the proportions shown in figure 19.

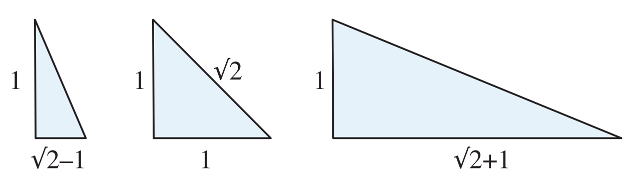


Figure 19. Proportions of triangles whose angles are multiples of 22.5°.

The origami design methodology known as tiling, described in [11–15], constructs crease patterns for complicated bases by fitting together simpler patterns that are composed of these triangles. These patterns commonly appear over and over at different scales. When all the creases run at multiples of 22.5°, the proportions of the squares, rectangles, and triangles that make up these patterns are all bilinear combinations of 1 and $\sqrt{2}$. Furthermore, the scaling factors that apply to these patterns are also such bilinear combinations. The upshot is that the dimensions in such a crease pattern are typically all related to each other by factors that are of the form $a + b\sqrt{2}$.

As an example, figure 20 shows one such crease pattern, used in an eagle that I designed some years back:

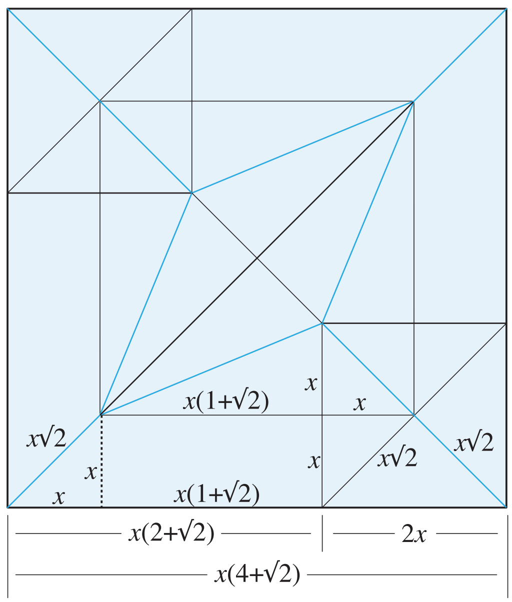


Figure 20. Crease pattern for the Eagle and relative proportions.

In this figure, I've marked in some of the proportions relative to a segment marked x. All of the segments are proportional to x. The proportions of adjacent triangles can be found by referring to the proportions of the three triangles shown in figure 2.

We can fill in the proportions of all segments until we get to the edge of the square; by summing the lengths of all segments along the edge, we find that the edge of the square is $x(4+\sqrt{2})$ units long. If one assumes a unit square, then

$$x = \frac{1}{4 + \sqrt{2}}. \tag{35}$$

To construct the origami crease pattern by folding, it is necessary to find the distance x—or any related distance, e.g., $x\sqrt{2}$, $2x$, or $x(1+\sqrt{2})$ — by folding. This could be done by several methods: a binary approximation or approximation as a rational by a continued fraction, followed by any of the rational methods (crossing diagonals, Fujimoto, Haga, or Noma).

It turns out, however, that many proportions of the form $a + b\sqrt{2}$, and this one in particular, can be folded exactly using a construction similar to the crossing-diagonals construction. Let's look again at the geometry of two crossing diagonals, shown in Figure 21.

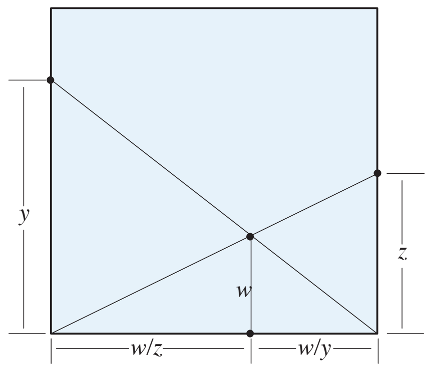


Figure 21. General form of the crossing-diagonals algorithm.

If the two diagonal creases hit the two sides at heights y and z, respectively, and we define w as the height of the intersection above the bottom of the paper, then dropping a line from the intersection divides the bottom of the square into segments of length $\frac{w}{z}$ and $\frac{w}{y}$, respectively. The total length of the bottom edge is thus

$$w\left(\frac{1}{y}+\frac{1}{z}\right). \tag{36}$$

Now, compare this form to the side length we computed based on the crease pattern in Figure 20, which was $x(4+\sqrt{2})$. If we equate the two, then we can seek to find an assignment of $w, x, y,$ and z that permits a relatively simple construction:

$$x\left(4+\sqrt{2}\right) \equiv w\left(\frac{1}{y}+\frac{1}{z}\right). \tag{37}$$

The simplest assignment is to take $x=w$. Then we are left with the equation

$$\left(4+\sqrt{2}\right) = \left(\frac{1}{y}+\frac{1}{z}\right). \tag{38}$$

If we could divide up $(4+\sqrt{2})$ into two pieces whose reciprocals are easy to find, then we'd have an exact solution for finding that particular division.

And as it turns out, there are many ways of performing this division. Let me first give a particular solution and show why it works, then I'll go back and explain other ways of doing it and give a general procedure.

The particular solution is:

$$(4+\sqrt{2}) = \left((2)+\left(2+\sqrt{2}\right)\right) = \left(\frac{1}{(1/2)}+\frac{1}{\left(1-1/\sqrt{2}\right)}\right), \tag{39}$$

so if we take $y = 1/2, z = 1 - 1/\sqrt{2}$, the crossing diagonals will divide the bottom of the paper as shown in Figure 21.

Finding $y=1/2$ is easy enough, but finding $z=1-1/\sqrt{2}$ is not immediately obvious. It turns out, though, that this proportion resides within the origami shape known as the Fish Base, as shown in Figure 22.

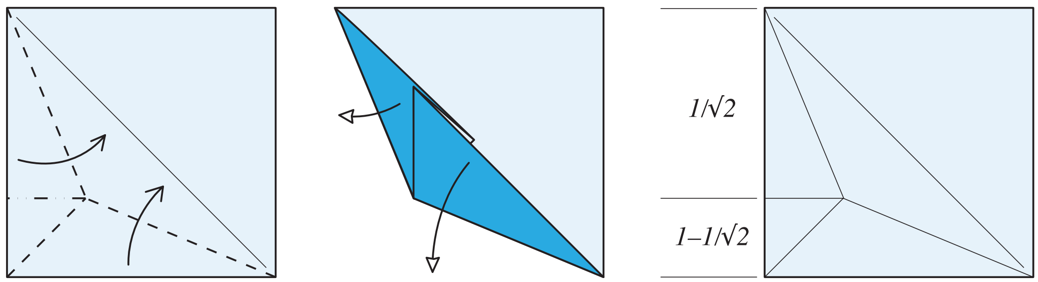


Figure 22. Construction of $1-1/\sqrt{2}$.

So if we start with a half Fish Base on one side and pinch a mark halfway up on the other, then the two crossing diagonals divide the bottom in the desired proportion, as shown in Figure 23.

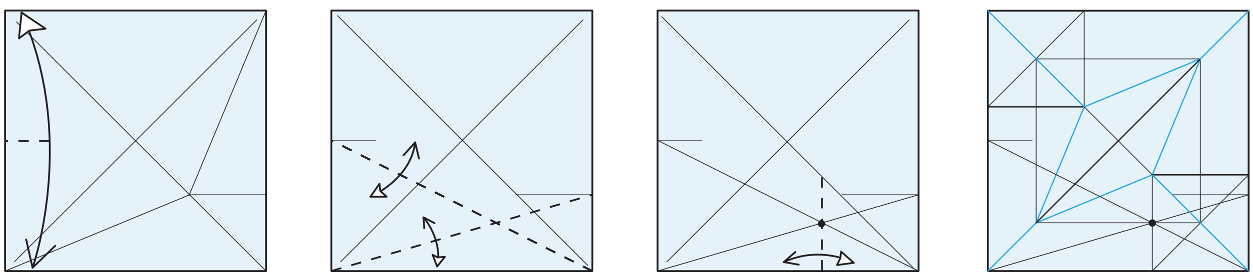


Figure 23. Folding sequence to find the initial division.

Essentially what we're doing is finding a reciprocal of the bottom edge by finding a division of the bottom edge in which the separate parts have easy-to-find reciprocals. In general, when the side of the square is of the form $x(a + b\sqrt{2})$, where x is the length of a significant crease in the pattern and a and b are rationals, one can usually find a crossing-diagonals sequence that gives the ratio x. Finding this sequence is tantamount to finding the reciprocal of $(a + b\sqrt{2})$. The trick to finding the crossing-diagonals sequence is to break up $(a + b\sqrt{2})$ into two terms for which we can easily find their reciprocals.

The integer or rational part a is usually not a problem, since we can find the reciprocal of any integer using the rational fraction constructions given earlier. The difficulty comes in identifying an easily foldable fraction whose reciprocal contains a term $b\sqrt{2}$.

Fortunately, there aren't too many of these and we can easily enumerate the most common possibilities. All are found by kite-folding, folding angles of 22.5°. Figure 24 shows the distance y, its reciprocal, and the creases that specify the desired proportion. The dashed line traces the associated diagonal crease, which would be one of a pair.

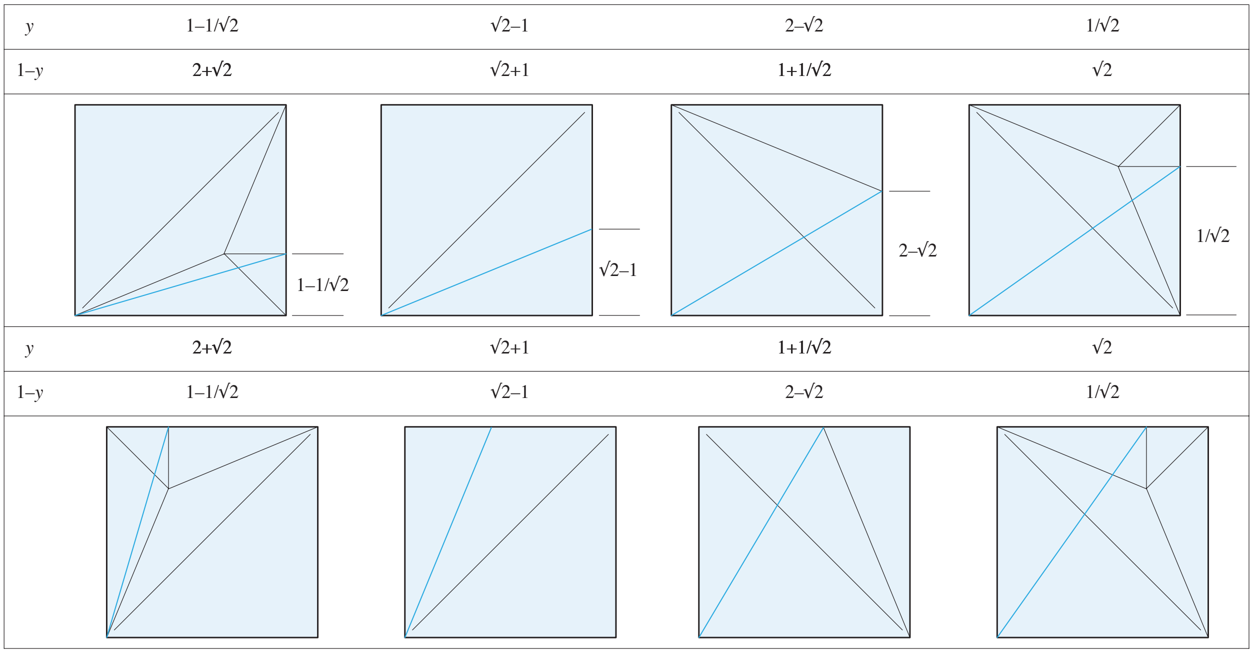


Figure 24. Common quadratic surds in origami, their reciprocals, and how to fold lines with slope equal to the value of the quadratic surd.

These tables give values of $1/y$ that contain factors $\pm\sqrt{2}$; but what about larger multiples of $\sqrt{2}$? That's easy; if you divide the fraction y by a factor b before forming the diagonal, the resulting reciprocal is *increased* by the same factor.

So the algorithm for finding the reciprocal of $(a + b\sqrt{2})$ is to let one diagonal give you the portion containing $\sqrt{2}$, and let the other diagonal give you the integer or rational portion. As with the purely rational constructions of the earlier sections, there are many possible ways to find the same proportion.

Angle Divisions

Less common than divisions of a line are divisions of angles; dividing an angle into thirds, fifths, or sevenths. Like divisions of a line, divisions of angles into powers of 2 are relatively easy. One might think that since division of a line into an arbitrary proportion is straightforward, simple solutions would exist for division of an angle into arbitrary proportions as well. But divisions of angles into other fractions are considerably harder.

In fact, it's well-known that using compass and straightedge, while a line segment can be divided into any number of equal divisions, division of an arbitrary angle into something as simple as thirds is impossible. Compass-and-straightedge construction is an ancient branch of mathematics — historical texts on the subject date back over two millennia. Solutions to compass-and-straightedge constructions give us many of the tools used in origami constructions, so let us digress for a moment to consider the mathematical field.

Many people encounter compass-and-straightedge problems in high school geometry. Compass-and-straightedge construction is similar to origami in several ways. In both, you are trying to produce geometric shapes, and both have stringent rules. In origami, of course, you use folding with no cutting. In compass-and-straightedge, you may use a compass, which is a tool for drawing circles, and an unmarked straightedge for drawing straight lines. It is a common part of

the elementary education to learn various geometric constructions: drawing a line through a point parallel to a given line, bisecting an angle, or drawing geometric figures such as an equilateral triangle, isosceles right triangle, or square. The roots of the field stretch back into antiquity; solutions for many constructions were described in Euclid's *Elements*, which was published sometime around the year 300 BCE.

Although many compass-and-straightedge constructions were devised by the ancients, there were three famous mathematical problems of antiquity that date back to the glory days of Greek mathematics in Athens some four hundred years BCE. and that have a special significance to origamists. The earliest great conundrum for which we have records was the problem of "squaring the circle," or constructing a square with the same area as a circle using compass and straightedge alone. The second was "doubling the cube," also called the "Delian problem" because it was attributed to the Apollonian oracle at Delos; the object is to construct the side of a cube whose volume is precisely double that of a given cube, or equivalently, given a line segment, construct a second segment that is exactly $\sqrt[3]{2}$ times as long. The third great problem, which is our interest here, was trisection of an arbitrary angle. Much of Greek mathematics (and in fact a substantial portion of modern mathematics) was devoted to the solution of these three problems. While an enormous body of mathematics grew out of this pursuit, it was all in vain, for ultimately all three compass-and-straightedge constructions were proven impossible some 2200 years later. While compass and straightedge allow one to draw both circles and lines, in origami, one can only fold straight lines. Thus it is rather surprising that angle trisection (and cube doubling, too, as it turns out) can be solved by origami techniques!

The advantage that origami has over compass and straightedge lies in the character of the numbers constructible by both. All numbers constructible by compass and straightedge can be written in terms of solutions of a quadratic equation, an equation in which the exponent of the unknown is no larger than 2. Given a set of lines of set length, one can with compass and straightedge construct any linear combination, multiple, or square root of those lengths. Thus with compass and straightedge, one can solve any quadratic equation or higher order equation that is reducible to quadratic equations whose coefficients are given as constructible distances.

However, the construction of the cube root of two and trisection of an arbitrary angle requires the solution of a cubic equation, in which the exponent of the unknown is 3, while squaring of the circle requires the construction of a segment of length π, which is a transcendental number that cannot be written as the root of a polynomial equation with less than an infinite number of terms. These three classical problems were proven impossible some 200 years ago.

A "proof of the impossible" of a different sort was a 1995 article in *The American Mathematical Monthly*, titled "Totally Real Origami and Impossible Paper Folding," in which the authors claimed to show that it was impossible to duplicate the cube using origami techniques [16, 17]. In fact, they claimed that origami was actually more restrictive than compass-and-straightedge constructions, and could not, for example, construct certain numbers of the form $\sqrt{1+\sqrt{2}}$ that are constructible by compass and straightedge.

In fact, solutions for duplication of the cube, trisection of an angle, as well as constructions of $\sqrt{1+\sqrt{2}}$ and related numbers have been known for many years in origami. The advantage of origami over compass-and-straightedge construction is that origami permits one to simultaneously align two separate points onto two different lines. The authors of the *Monthly* article considered a subset of the known origami operations that did not allow this type of simultaneous alignment. However, the simultaneous alignment of two points onto two lines permits the solution of cubic equations and therefore, solution of two of the classical problems of antiquity: duplication of the cube and trisection of a given angle.

Therefore origami *can* solve cubic equations, and since angle trisection requires solution of a cubic equation, it would appear that origami could also trisect an arbitrary angle — the second classical problem. Indeed it can, and there are several such constructions. One solution for trisecting an acute angle in the corner of a square, devised by the Japanese folder and mathematician Tsune Abe [18, 19], is illustrated in figure 25.

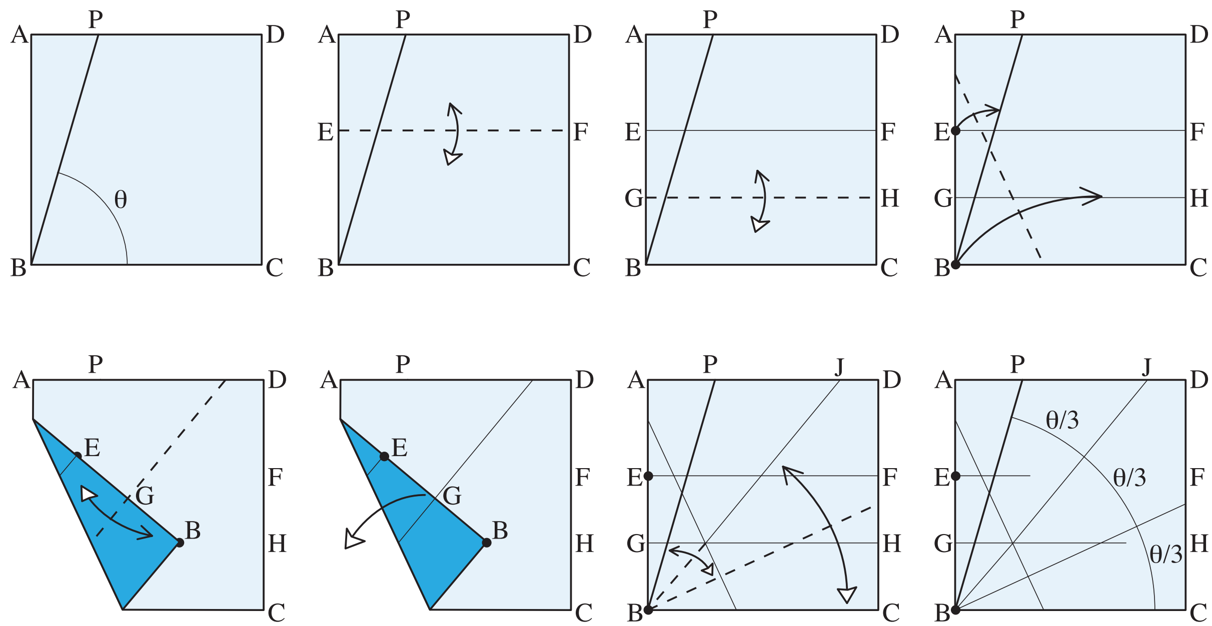


Figure 25. Tsune Abe's trisection of an arbitrary acute angle.

The procedure for Abe's trisection is the following:

(1) Mark the angle to be trisected inside one corner of the square. In this example, angle PBC is to be trisected.

(2) Fold any crease parallel to edge BC.

(3) Fold edge BC up to crease EF and unfold.

(4) Fold corner B up so that point E lies on line BP and corner B lies on line GH.

(5) Crease along an existing crease through point G, creasing through all layers.

(6) Unfold.

(7) Extend the crease from point J back to point B. Also, bring edge BC to fold BJ and unfold.

(8) The angle is trisected.

A technique for trisecting obtuse angles devised by the French folder and mathematician Jacques Justin, is illustrated as well in figure 2 [20]. (Since any angle can be trisected by trisecting its complement, either technique can be used for any angle.) Justin's technique does not require use of the corner of the square and is illustrated as if in the middle of an infinite sheet. The key

observation to note is that both techniques require the simultaneous alignment of two points on a line.

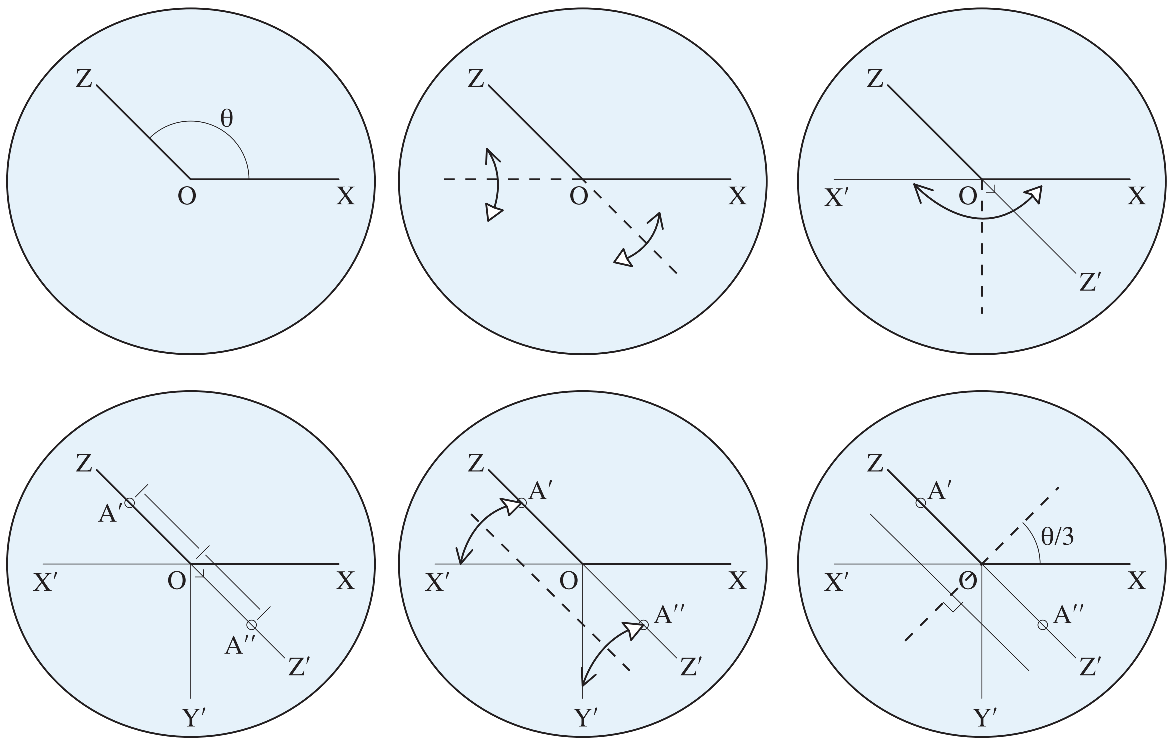


Figure 26. Jacques Justin's trisection of an obtuse angle.

Justin's trisection is the following:

(1) The angle to be trisected is angle ZOX.

(2) Extend lines ZO and XO.

(3) Fold X to X′ through point O.

(4) Mark off points A′ and A″ on lines ZO and Z′O at equal distances from point O.

(5) Fold points A′ and A″ to lie on lines X′O and Y′O and unfold.

(6) Fold a line perpendicular to the last crease through point O to trisect the angle.

Angle trisection and bisection can be combined to divide the unit circle into many different divisions, or equivalently, to construct a regular polygon of *N* sides (a "regular *N*-gon"), where *N* is of the form $2^n 3^m$ (*n* and *m* are arbitrary integers). Thus, using only folding, one can divide any angle into equal divisions numbering 2, 3, 4, 6, 8, 9, 12, and so forth.

For the particular case where you are dividing a complete circle into *N* equal parts, there is another family of origami constructions discovered by the Austrian mathematician Robert Geretschläger [21–24], based on geometric constructions dating back to the 1890s [25]. He has shown a general approach for constructing a regular *N*-gon where *N* is a prime number of the form $2^n 3^m + 1$. The numbers of this form are 3, 5, 7, 13, 17… This construction can be combined

with angle bisection and trisection as well to give other polygons of the form $2^j 3^k \left(2^n 3^m + 1\right)$ whenever the term in parentheses is prime. Although a full description of Geretschläger's approach is well beyond the scope of this article, the references at the end of this section illustrate several specific cases and the general approach. Using these constructions, the only nonconstructible regular N-gon for $N \leq 20$ is $N = 11$.

Exact constructions of angular divisions are *tours de force* of mathematics, but they are usually impractically complex to be used for origami design, in that they cover the paper with incidental creases and can require inherently inaccurate creasing: long narrow triangles, distant extrapolations using creases, copying of angles and distances.

However, as we have seen with divisions of an edge, for practical purposes, an approximation can often be as good or better than an exact solution. In fact, we can use edge division to construct approximations to angular divisions.

An example from my own work will illustrate this process. In my book, *The Complete Book of Origami*, a Scorpion design required division of a 90-degree angle into sevenths in the early stages of the model [26]. This is not terribly difficult to find by trial-and-error (fan-fold the angle into sevenths and continuously adjust the creases until all divisions are equal), but we can also find an approximate solution that is deterministic and is accurate to within folding error.

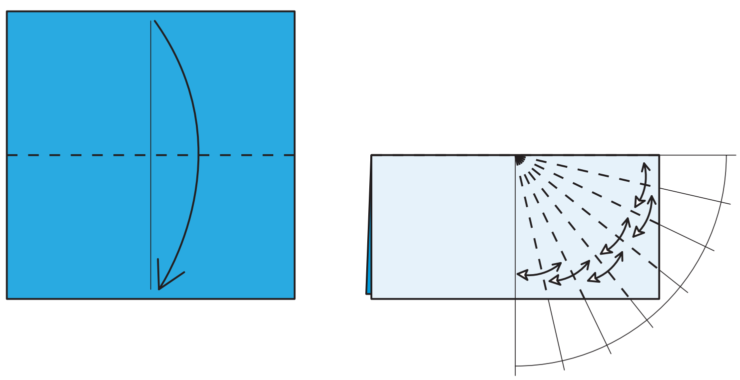


Figure 27. First 2 steps of Lang's Scorpion, which entails a division of an angle into sevenths.

Now, we could approach this two ways: we could try to divide the angle itself into sevenths, or we could try to locate the points on the edge of the paper where one or more of the creases hits the edge of the paper. If we're clever about this, we'll only have to locate one of them; if, for example, we found the line for 4/7 of the angle, we could then bisect it twice to get 2/7 and 1/7, and subsequently all the other divisions, purely by folding.

Now there is no simple algebraic expression for these points' locations, but using some high-school trigonometry, we can calculate where the creases hit the edge; the decimal values of the numbers are shown on an unfolded square in figure 2. The distances, expressed as a fraction of the edge of the square, are given by the formula

$$y_i = \frac{1}{2}\left(1 + \tan\frac{90°}{7} i\right), \qquad (40)$$

where i is the index of the angle shown in Figure 28.

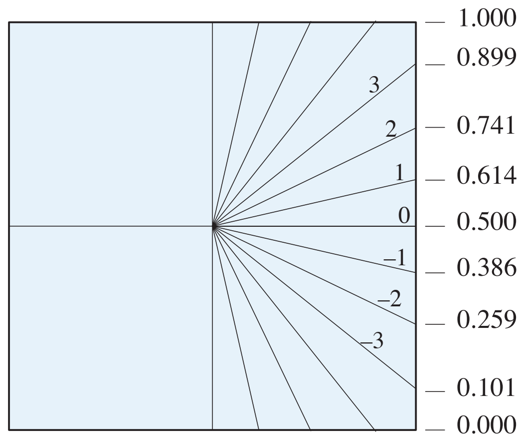


Figure 28. Intersections of seventh angular divisions with the edge of the paper.

Any one of these could be approximated by the binary method or by a rational fraction derived from the convergents of the continued fraction. Noting that $y_1 = 0.101 \approx 1/10$ leads to the folding sequence shown in Figure 29.

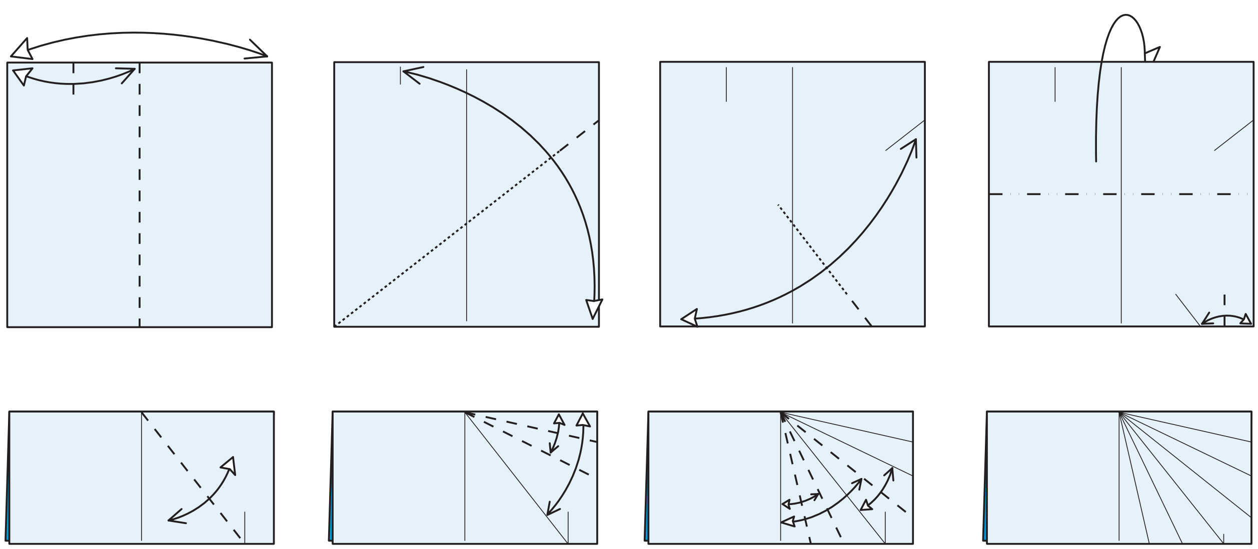


Figure 29. Folding sequence for dividing the central 90° angle into sevenths.

It is also possible to use an iterative approximation to any angular division, based on the binary method, employing successive bisection of the angle (just as the binary method employed successive division). If we equate the rays on either side of an angle with the top and bottom edges of the square, then there is a natural correspondence between the folds that divide the edge of the square and the folds that divide an angle, as shown in Figure 30.

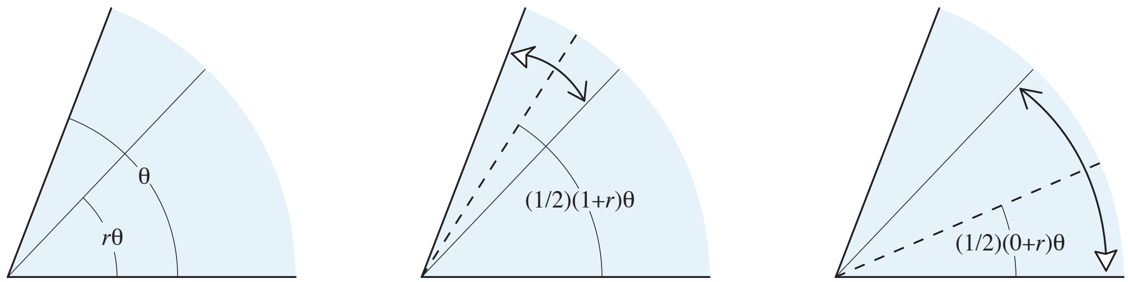


Figure 30. Division of an angle by bisection corresponds to the two operations that make up the binary folding method.

If we use the two operations shown in Figure 30, then we can apply these two operations according to the binary expansion of a fraction r to divide the angle in the ratio r:1–r. For non-binary fractions (like 1/3), the infinite but repeating binary expression for the fraction gives an iterative method of division. Thus, for example, dividing the angle into 7ths, which has the binary expansion

$$\frac{1}{7} = .\overline{001}, \tag{41}$$

can be accomplished by repeating the procedure (left, left, right), where "left" and "right" refer to the two sides of the angle to be divided into 7ths.

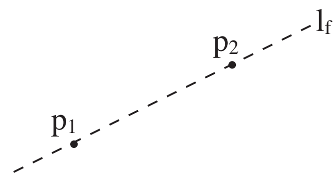
Axiomatic Origami

The folding methods I've shown thus far use the same basic operations in different combinations: fold a point to another point, fold a line to another line (angle bisection), put a crease through one or two points. Starting in the 1970s, several folders began to systematically enumerate the possible combinations of folds and to study what types of distances were constructible by combining them in various ways. The first systematic study was by Humiaki Huzita [27–29], who described a set of six basic ways of defining a single fold by aligning various combinations of existing points, lines, and the fold line itself. These six operations have become known as "Huzita's Axioms" (HA), although they may be best thought of as operations that act upon points and lines. Given a set of points and lines on a sheet of paper, Huzita's operations allow one to create new lines; the intersections among old and new lines define additional points. The expanded set of points and lines may then be further expanded by repeated application of the operations to obtain further combinations of points and lines.

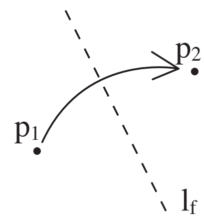
The set of points constructible by repeated application of HA to some initial set of features—typically, the corners and edges of the unit square—are of both academic and practical interest. From the academic side, it has been shown that HA can be used to construct distances that are solutions to cubic equations by sequential single folds. In particular, elegant constructions have been presented for two of the three great problems of classical antiquity that are not possible with compass and unmarked straightedge: angle trisection, as we have seen, and doubling of the cube [30], which we will shortly encounter.. On the practical side, HA can give both exact and approximate folding sequences of very low rank.

A particularly clear and lucid account of HA is given at [31]. Although called "axioms" they are best thought of as fundamental operations that act on points and lines to produce a new line, which is the fold line. The six operations identified by Huzita are shown in Figure 31.

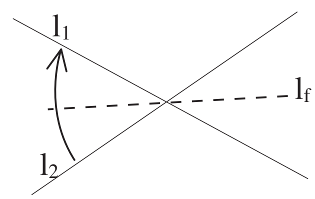
(O1) Given two points p_1 and p_2, we can fold a line connecting them.



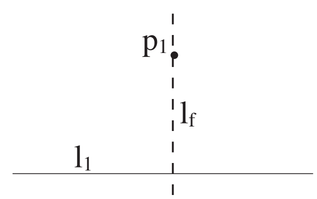
(O2) Given two points p_1 and p_2, we can fold p_1 onto p_2.



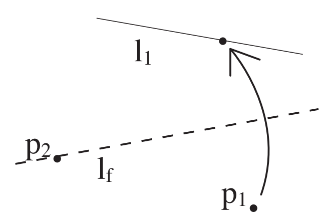
(O3) Given two lines l_1 and l_2, we can fold line l_1 onto l_2.



(O4) Given as point p_1 and a line l_1, we can make a fold perpendicular to l_1 passing through the point p_1.



(O5) Given two points p_1 and p_2 and a line l_1, we can make a fold that places p_1 onto l_1 and passes through the point p_2.



(O6) Given two points p_1 and p_2 and two lines l_1 and l_2, we can make a fold that places p_1 onto line l_1 and places p_2 onto line l_2.

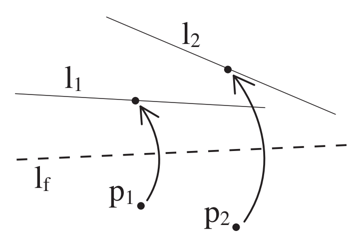


Figure 31. The six operations of Huzita's Axioms.

As we will see, operations O1–O5 can be used to construct the solution of any quadratic equation with rational coefficients. Operation O6 is unique in that it allows the construction of solutions to the general cubic equation.

Recently, a 7th operation was proposed by Hatori [32], which I will denote by (O7). It is shown in Figure 32.

(O7) Given a points p_1 and two lines l_1 and l_2, we can make a fold perpendicular to l_2 that places p_1 onto line l_1.

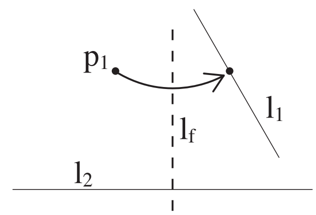


Figure 32. Hatori's 7th axiom.

Hatori noted that this operation was not equivalent to any of HA. Hatori's O7 allows the solution of certain quadratic equations (equivalently, it can be constructed by compass and straightedge). If we denote the expanded set as the "Huzita-Hatori Axioms" (HHA), it turns out that this set is complete; these are all of the operations that define a single fold by alignment of points with finite line segments. Over the next section, I will show that this set is complete.

Preliminaries

The proof of completeness and enumeration relies in part on counting degrees of freedom in a system of operations. This enumeration is aided by creating an algebraic description of points, lines, and operations.

Definition: a *point P* is an ordered pair (x,y) in $\Re^2$ with $x \in [-\infty,\infty]$, $y \in [-\infty,\infty]$.

We note that a point has 2 *degrees of freedom* (DOF), i.e., two parameters that can be varied independently, namely, the two coordinate values.

Lines are a bit more complicated; a line can be defined in several ways. One possibility proceeds from O1, which corresponds to one of Euclid's axioms: "through any two points there exists exactly one line." This suggests that a line be defined by two different points somewhere upon it. Since each point is defined by two coordinates, that definition would require that four coordinate values be used to define any line. However, such a definition is not unique; one could define the same line by any two pairs of points.

A second, more parsimonious definition is suggested by the high-school algebra equation of a line in Cartesian coordinates: $y = mx + b$, where m is the slope and b is the y-intercept, and the line is defined as all coordinate pairs (x,y) that satisfy this equation. This expression makes it clear that a line, too, has 2 DOF; the two coordinate values m and b are sufficient to uniquely describe nearly any line.

A deficiency of using the Cartesian equation is that it does not uniquely specify lines parallel to the y-axis (which have infinite slope m and the intercept b is undefined). It is more useful to adopt a parameterization that does not require infinite values and that treats all lines in some sense "equally."

I find it useful to characterize a line by a 2-vector perpendicular to the line and a particular point on the line, according to the following.

Definition: Define the directed unit vector $U(\alpha)$, as

$$U(\alpha) \equiv (\cos\alpha, \sin\alpha) \text{ for any } \alpha \in [0, 180°). \tag{42}$$

Definition: A line $L(d,\alpha)$ is the set of all points *P* that satisfy the equation

$$(P - dU(\alpha)) \cdot U(\alpha) = 0, \tag{43}$$

for any $d \in [-\infty, \infty]$, $\alpha \in [0, 180°)$, and $A \cdot B$ denotes the scalar product of *A* and *B*. It is not hard to show that with this definition, any line is specified by a unique combination (d, α). It is also easy to show that equation (43) is equivalent to

$$P \cdot U(\alpha) - d = 0. \tag{44}$$

A convenient parameterization of the line $L(d,\alpha)$ is given by the following.

Definition: Given a vector $P = (x, y)$, the perpendicular vector $P^{\perp}$ is defined as

$$P^{\perp} \equiv (y, -x). \tag{45}$$

$P^{\perp}$ is P having undergone a 90° counterclockwise rotation. As a point of simplified notation, I will define $U^{\perp}(\alpha) \equiv (U(\alpha))^{\perp}$.

⊥Then it is easily shown that every point *P* on the line $L(d,\alpha)$ can be expressed in the form

$$P = dU(\alpha) + tU^{\perp}(\alpha) \text{ for some } t \in [-\infty, \infty]. \tag{46}$$

The geometric interpretation of equation (46) is shown in Figure 33. The point $dU(\alpha)$ is the point on the line closest to the origin; the offset $tU^{\perp}(\alpha)$ shifts the point $dU(\alpha)$ along the line by a distance *t*, which can be either positive or negative.

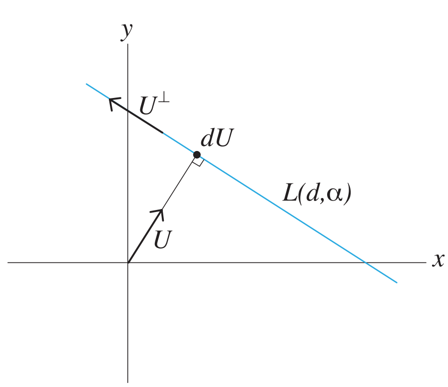


Figure 33. Geometric interpretation of the parameterization in equation (4).

Every point of the form (4) satisfies equation (2) and vice-versa; thus, either equation may be used as the definition of a line.

Folding

A fold is defined by a line, called the *fold line*. The fold line divides the paper into two regions. On one side of the line is the stationary region; the other side is the moving region. The choice of which is stationary and which is moving is completely arbitrary and the names serve only to aid intuition.

When a fold is formed, all features in the moving region have their coordinates reflected through the fold line, which will be denoted by $L_F(d_F, \alpha_F)$.

Since a fold is defined by a line, and a line has two DOF (namely, the parameters d_F and α_F), it takes two DOF to fully specify the fold line.

For notational simplicity in what follows, I will define $U_F \equiv U(\alpha_F)$.

If the fold line is given by $L_F(d_F, \alpha_F)$, then a point P within the moving region is, after the fold, located at a point P' given by

$$\begin{aligned} P' &= P - 2\big((P - dU_F)\cdot U_F\big)U_F \\ &= P + 2\big(d_F - P\cdot U_F\big)U_F \end{aligned}. \tag{47}$$

We will denote the result of folding a point *P* by *F(P)*. That is,

$$F(P) \equiv P + 2\big(d_F - P\cdot U_F\big)U_F. \tag{48}$$

It is relatively straightforward to verify two identities:

$$\text{For any point } P,\ F(F(P)) = P, \tag{49}$$

which simply states the obvious fact that folding a point back and forth along the same fold line leaves it unchanged.

For any point *P* on the fold line $L_F(d_F, \alpha_F)$,

$$F(P) = P, \tag{50}$$

which states that a point on the fold line is unchanged by a fold.

We will also define the result of a fold acting on a line. For a line *L*, we denote by *F(L)* the set of all points *F(P)* such that (a) *P* satisfies equation (44), (b) *P* lies within the moving region of the paper.

Alignments

We now describe what it means to bring two features into alignment using a fold. A single alignment consists of bringing two features together. We have two types of features: points and lines. We must first define what we mean by alignment.

Bringing a point to a point $P \leftrightarrow P$

Two points $P_1 \equiv (x_1, y_1)$ and $P_2 \equiv (x_2, y_2)$ are said to be *aligned* when both their coordinate values are equal. We denote alignment by a double-headed arrow: $P_1 \leftrightarrow P_2$. That is,

$P_1 \leftrightarrow P_2$ if and only if $x_1 = x_2$ and $y_1 = y_2$. (51)

Since two equations must be satisfied, aligning two points consumes two DOF.

Bringing a point onto a line ($P \leftrightarrow L$ **)**

A point $P_1 \equiv (x_1, y_1)$ is said to be aligned with a line $L(d,\alpha)$ if and only if it lies on the line, that is, if P_1 satisfies equation (44). We denote alignment between and point and a line by the same double-headed arrow: $P_1 \leftrightarrow L(d,\alpha)$. That is,

$P_1 \leftrightarrow L(d,\alpha)$ if and only if $P_1 \cdot U(\alpha) - d = 0$. (52)

Since only one equation must be satisfied, aligning a point to a line consumes only one DOF.

We note that the alignment operator is defined to be commutative; that is, for dissimilar operands,

$P_1 \leftrightarrow L(d,\alpha)$ if and only if $L(d,\alpha) \leftrightarrow P_1$. (53)

Bringing one line to another line ($L \leftrightarrow L$ **)**

Two lines $L_1(d_1,\alpha_1)$ and $L_2(d_2,\alpha_2)$ are said to be aligned if and only if every point in L_1 is aligned with L_2 and vice-versa.

For simplicity of notation, let us denote $U_1 \equiv U(\alpha_1)$, $U_2 \equiv U(\alpha_2)$. Then if we choose the parameterization of equation (46) to define line L_1, that is, a point P_1 on line L_1 is given by

$$P_1 = dU_1 + tU_1^{\perp}, \tag{54}$$

for some (d,t), then alignment of the two lines implies that every such point P_1 must satisfy equation (44), namely:

$$\left(d_1 U_1 + tU_1^{\perp}\right) \cdot U_2 - d_2 = 0. \tag{55}$$

A bit of rearranging gives

$$\left(d_1\left(U_1 \cdot U_2\right) - d_2\right) + t\left(U_1^{\perp} \cdot U_2\right) = 0. \tag{56}$$

The left side of equation (56) is linear in *t*; for the equation to be satisfied for all *t*, both the linear term and the constant must individually be equal to zero. Thus:

$$d_1\left(U_1 \cdot U_2\right) - d_2 = 0, \tag{57}$$

$$U_1^{\perp} \cdot U_2 = 0. \tag{58}$$

Consequently, for two separate lines to be brought into alignment, two equations must be satisfied and two DOF are consumed.

A geometric interpretation of the two equations is that equation (58) enforces that the two lines are parallel, while equation (57) enforces that they intersect.

In fact, it can easily be shown that if the two lines are known to have a point of intersection, then equation (58) is sufficient.

Alignments by folding

In the previous section, I defined the three basic types of alignments: $P \leftrightarrow P$, $P \leftrightarrow L$, $L \leftrightarrow L$. I will now enumerate all possible alignments that may be created by a single fold. Such alignments may be made between preexisting features on the paper, or may include the feature created by the fold, namely, the fold line itself.

We consider (and dismiss) alignments between two preexisting features that are both moving or both stationary. Any such alignments are not created by the fold and thus cannot be used to specify the location of the fold line. The remaining, interesting classes of alignments are those between two preexisting features where one is moving and one is stationary, and alignments between a preexisting feature and the fold line.

Consider first alignments between two features that already exist on the paper, one of which is on the moving portion and the other must be on the stationary portion. This gives rise to 5 possible alignments, which are given in Table 9.

Symbol	Description	# of Equations
$F(P_1) \leftrightarrow P_2$	Fold point P_1 to another point P_2	2
$F(P_1) \leftrightarrow L$	Fold point P_1 to line L	1
$F(L) \leftrightarrow P$	Fold line L to point P	1
$F(L_1) \leftrightarrow L_2$	Fold line L_1 to different line L_2	2
$F(L) \leftrightarrow L$	Fold line L onto itself	1

Table 9. The five distinct nontrivial alignments between points and lines.

We must distinguish the last two cases, because while folding a line onto another line requires the solution of two conditions (equations (15) and (16)), when folding a line onto itself, the line and its image under folding intersect at the fold line; thus it is sufficient to require only equation (16).

The second set of alignments consists of alignments between preexisting features and the fold line. There are two possibilities: aligning a point with the fold line, and aligning a line with the fold line. The latter case is trivial; making the fold along an existing line creates no new features. So the only nontrivial case is aligning a point with the folding line, given in Table 10.

Symbol	Description	# of Equations
$P \leftrightarrow L_F$	Align point P with the fold line L_F	1

Table 10. The sole distinct nontrivial alignment between a point and a fold line.

This completes the listing of all nontrivial alignments that can be created by a single fold.

Multiple Alignments

Now, we would like to use alignments to define the fold; that is, by specifying one or more alignments, we completely specify the location of the fold line (or equivalently, its two parameters d_F and α_F). This requires that we have as many equations created by the alignments as we have unknowns: two. We observe that there are two alignments that each by itself imposes two equations. They are $F(P_1) \Leftrightarrow P_2$ (fold one point to another point), and $F(L_1) \Leftrightarrow L_2$ (fold one line to another line). These two alignments are individually sufficient to define a fold line; they correspond to Huzita's axioms O2 and O3, respectively, and are given in Table 11.

$F(P_1) \Leftrightarrow P_2$	O2
$F(L_1) \Leftrightarrow L_2$	O3

Table 11. The two operations that specify two DOF and the HA that they correspond to.

The other four alignments only create a single equation; we must therefore take pairs of them to create two equations to fully specify the fold line. With four possible alignments, there are 10 possible distinct pairs (since the order is unimportant), which are summarized in Table 12.

	$F(P_2) \Leftrightarrow L_2$	$F(L_2) \Leftrightarrow P_2$	$F(L_2) \Leftrightarrow L_2$	$P_2 \Leftrightarrow L_F$
$F(P_1) \Leftrightarrow L_1$	O6			
$F(L_1) \Leftrightarrow P_1$	O6	O6		
$F(L_1) \Leftrightarrow L_1$	O7	O7	N/P	
$P_1 \Leftrightarrow L_F$	O5	O5	O4	O1

Table 12. Possible alignment pairs that specify a single fold and their corresponding HHA.

One combination, $(F(L_1) \Leftrightarrow L_1, F(L_2) \Leftrightarrow L_2)$ has no solutions if L_1 and L_2 are nonparallel and infinite solutions if they are parallel. Each of the remaining pairs correspond to one of the Huzita-Hatori axioms. Since these represent all possible alignments that create exactly two degrees of freedom, this shows that the HHA set is complete (and that Hatori's 7th axiom is indeed necessary for completeness).

Constructability

It is relatively straightforward to construct explicit expressions for the fold line parameters (d_F, α_F) for six of the seven HHA operations in terms of the parameters of the constituent points and lines. Each involves the solution of equations no more complicated than quadratic, and indeed, the six operations can be used to construct exact solutions for any quadratic equation with rational coefficients.

However, operation O6—fold two points to two lines—is more complicated. An analytic solution for the fold line requires that one solve a cubic equation, which means that by performing this maneuver, one can solve cubic equations exactly.

Perhaps the most famous example of solving a cubic using this fold is Peter Messer's solution for the doubling of the cube—or more specifically, constructing two segments whose lengths are in the proportion of $\sqrt[3]{2}$. This beautiful construction was presented in [30], and is reproduced in Figure 34. The square is divided into thirds by horizontal creases. Then the corners is folded so

that points P_1 and P_2 lie on lines L_1 (the left edge) and L_2 (the upper horizontal crease). The point where P_1 hits the edge divides it in the desired proportion.

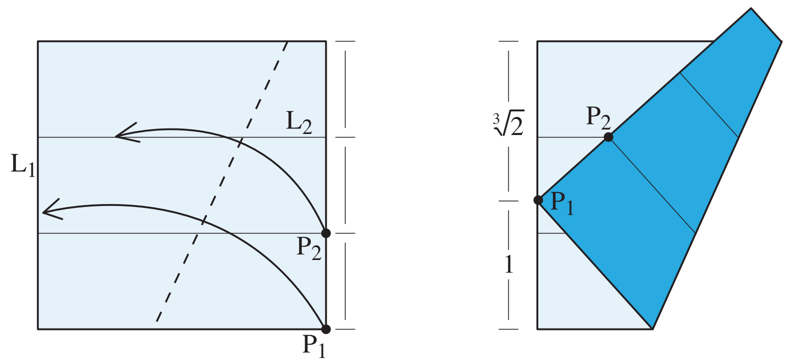


Figure 34. Peter Messer's construction of $\sqrt[3]{2}$.

In fact, Axiom O6 has connections to several interesting branches of mathematics, and so is worth a bit closer study.

Axiom 6 and Cubic Curves

Given any two points and two lines on a sheet of paper, Axiom 6 states that it is possible to fold both points onto both lines. This is an incomplete generalization; as it turns out, it is not always possible for some combinations of points and lines and for others it is possible in more than one way. We should consider all possible combinations of two points P_1, P_2 and two lines L_1, L_2. To simplify the examination, we will adjust our coordinate system so that one of the lines, L_1, is the x-axis, so that

$$d_1 = 0,\ U_1 = (0,1). \tag{59}$$

Similarly, we can assume with no loss of generality that point P_1 is located at

$$P_1 = (-1,0). \tag{60}$$

We also assume that line L_1 is stationary and point P_1 is moving.

We now assume a fold line L_f, characterized by parameters d_f and a_f such that

$$U_f = \left(a_f, \sqrt{1-a_f^2}\right). \tag{61}$$

We define P_1' as the image of P_1 under the fold L_f, that is,

$$\begin{aligned} P_1' &= F(P_1) \\ &= P_1 + 2(d_F - P_1 \cdot U_F)U_F \end{aligned} \tag{62}$$

If we require that fold line L_f places point P_1' onto line L_1, then equation (44) must be satisfied, i.e.,

$$P_1' \cdot U_1 - d_1 = 0. \tag{63}$$

Substituting (59–62) into (63) and solving for d_f gives

$$d_f = \frac{2a_f^2 - 1}{2\sqrt{1 - a_f^2}}, \tag{64}$$

leaving only a single free parameter (a_f) to specify the fold line.

For any fold line parameter a_f, any point P_2 has an image P_2' that results from the action of folding about the fold line. For a given point P_2, as we vary a_f over its range from –1 to 1, the point P_2' sweeps out a curve in space. Two such curves, for the points $P_2 = (2,-1.5)$ and $P_2 = (2,+1.5)$, are illustrated in Figure 35.

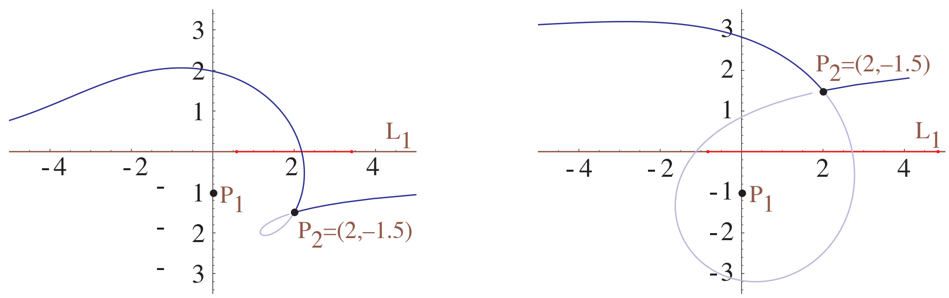


Figure 35. Locus of points swept out by P_2 as the fold line parameter a_f varies from –1 to 1 for two points P_2.

The question of whether there is a solution that places P_2' on a given line L_2 is answered by whether, and how many times, line L_2 intersects the curve swept out by P_2'. Thus, it is useful to derive an equation for this curve.

If we define

$$P_2 \equiv (a,b), \tag{65}$$

$$P_2' \equiv (x,y), \tag{66}$$

and solve for the coordinates of P_2', we find that

$$x = a - 2aa_f^2 + \frac{a_f\left(2a_f^2(1+b) - (1+2b)\right)}{\sqrt{1 - a_f^2}}, \tag{67}$$

$$y = 2aa_f\sqrt{1 - a_f^2} + \left(2a_f^2 - 1\right)(1+b). \tag{68}$$

Eliminating a_f from the two equations gives an equation for the shape of the curve swept out by P_2':

$$\begin{aligned} &y^3 + (1-b)y^2 + \left(x^2 - b(2+b) - a^2\right)y + \\ &\left(b^3 + b^2 + a^2b - a^2 + 2ax - x^2 - bx^2\right) = 0 \end{aligned}. \tag{69}$$

Equation (69) contains terms at most cubic in *y*, quadratic in *x*, with no term of degree higher than 3; it is a cubic curve, a type of curve that figures in many branches of mathematics.

Figure 35 illustrates several salient features of such a curve, which may be derived from equation (69).

- As $a_f \to \pm 1$, the curve approaches the asymptotes $(\pm\infty, 1+b)$.
- It usually contains a loop (colored light gray in Figure 35); the crossing of the loop occurs at point $P_2 = (a,b)$.
- Any line L_2 cuts the curve in at most 3 places; thus, there at most 3 possible alignments of P_2 onto L_2. If line L_2 only cuts the curve in 1 place, then there is only one possible alignment; and if L_2 misses the curve entirely, there are no possible alignments.

If we impose a change of variable on this curve,

$$\begin{aligned} x &\to x + a \\ y &\to y + b \end{aligned}, \tag{70}$$

the equation takes on the homogeneous form

$$y^3 + (1+2b)y^2 + y(2a+x)x = x^2, \tag{71}$$

which for the special case *b*=–1/2, is called an Ophiuride curve [33], and for *a*=0, *b*=–1/2, is called the Cissoid of Diocles [34].

I implicitly assumed in the analysis above that L_2 was stationary and P_2 was moving; to be consistent with this assumption, for each fold line L_2, we must insure that P_1 and P_2 both lie on the same side of the fold line. It can be shown that this condition holds everywhere along the curve except within the loop; thus, only the black portion of the curve corresponds to a physically realizable alignment.

We can also use this curve (and this alignment) to solve a general cubic equation. If we set equation (26) equal to a general cubic,

$$y^3 + ry^2 + sy + t, \tag{72}$$

and equate coefficients, we can find two solutions for *a*, *b*, and *x*, which turn out to be fairly complicated algebraic expressions but that involve only square roots. Taking the particular case

$$y^3 - 2 = 0, \tag{73}$$

we find two possible solutions:

$$P_2 = (a,b) = (\pm 1,1), \;\; x = \pm 2. \tag{74}$$

Consequently, one solution for $\sqrt[3]{2}$ is given by placing the point P_1 on L_1 and $P_2 = (1,1)$ onto vertical line $x = 2$, as illustrated in Figure 36, with the other solution given by its mirror image. The y-coordinate of P_2' gives the desired proportion.

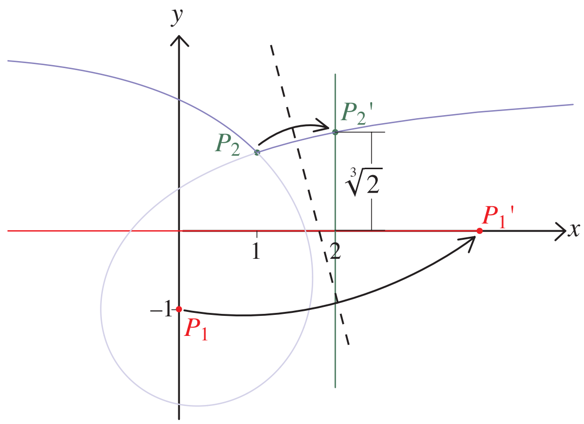


Figure 36. Another folding sequence to construct $\sqrt[3]{2}$.

By using operation O6, it is possible to solve cubic equations. What about higher-order equations? As noted already, Robert Geretschläger has shown how to construct regular *N*-gons when *N* is a prime number of the form $2^n 3^m + 1$, a "Pierpont prime" [25]. Such a construction is tantamount to solving the cyclotomic equation,

$$z^N - 1 = 0, \tag{75}$$

for its complex roots *z*. It is not know whether the Pierpont primes go on forever—there are 42 such primes below 1,000,000—but clearly, solving equation (30)—which can be done with folding—provides examples of solving high-order polynomial equations. However, the solution relies on the fact that for Pierpont primes, equation (30) can be factored in a way that requires solution of only cubic and quadratic equations. The more general question is whether an *irreducible* higher-order equation is solvable by folding.

If we require that all folds occur one at a time, then the seven HHA operations define all possible alignments, and since they collectively solve only quadratic and cubic equations, the answer is "no." However, if we broaden the acceptable operations to include alignments that specify multiple folds, the answer appears to be that at least *some* irreducible higher-order polynomial equations are solvable by folding.

Consider, for example, the following operation. We make fold L_{f1} that moves point P_1 onto line L_1 and fold L_{f2} that moves point P_2 onto line L_2; we additionally require that L_{f1} moves a point P_3 and L_{f2} moves a point P_4 so that their images P_3' and P_4' are aligned with each other, as shown in Figure 37. The solutions to such an operation will be defined by the points of intersection between two cubic curves of the form of equation (26), which are overlaid on the figure. There are five possible points of intersection, corresponding to 5 possible solutions for the

two folds. The points of intersection are indicated by colored dots; from these working out the necessary folds is relatively straightforward. Without the aid of the two cubic curves, finding the five possible solutions by empirical trial and error is quite challenging.

It is possible to get at least 7 points of intersection between two such curves, indicating that the defining polynomial can be of at least order 7. While this argument does not address whether all such polynomials are irreducible, given the generality of the configuration, the prospect seems unlikely. The question is also quite open as to which general higher-order polynomials can be translated into a folding problem.

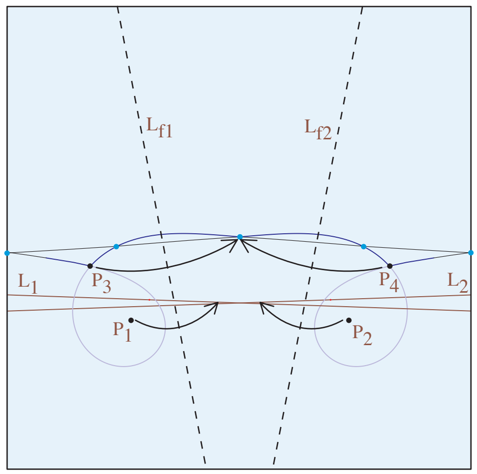


Figure 37. A two-fold construction whose solution is given by the intersections of two cubic curves.

Approximation by Computer

As we have seen, it is possible to approximate any proportion along the edge of a square to arbitrary accuracy. Indeed, it is similarly possible to locate any point in the interior of a square; all one needs to do is locate the *x*- and *y*-coordinates of the point along two adjacent sides, then project inward perpendicular creases, as Figure 38 shows for the point (3/8,5/8). The intersection of the creases that define the two coordinates gives the desired reference point.

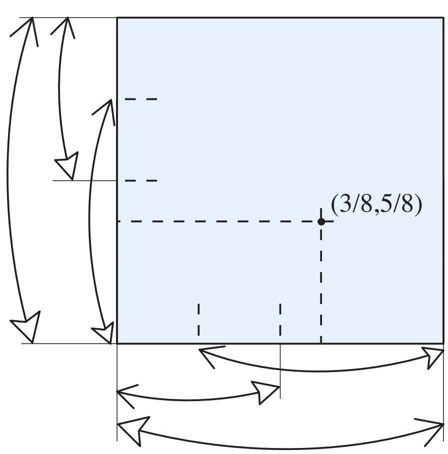


Figure 38. Construction of the point (3/8,5/8) in the interior of a square from reference points along the edges.

The points along the edges may be approximated in several ways, as we have seen. The simplest is the binary method used in the figure. Recall that with the binary method, approximating an arbitrary location along a single edge to an accuracy of .005 required at most 9 folds. To locate a point within the interior to this accuracy would then require 18 folds.

However, constructing the x and y coordinates independently is a rather inefficient method of locating a point. The binary algorithm for locating a point only makes use of one of the seven Huzita-Hatori operations, specifically O2 ("given two points p_1 and p_2, we can fold p_1 onto p_2"), and then only considers points along a single edge of the square. Might we do better by using any of the other operations? And if so, how?

Let us consider a broader question: Given a square, how many distinct points can we create using no more than r folds? If there are a large number of points constructible in relatively few steps, the odds are good that for any desired reference point, one of the constructible points is fairly close to the desired point. And this can be quantified; with N constructible points, so long as the points are distributed roughly uniformly, then for any desired reference point, there is a constructible point within a distance $N^{-1/2}$ of the target, on average. So, for example, with 10^6 constructible points, for any given target point, one of the constructible points is, on average, only about .001 units away, anywhere in the square.

The question then becomes: what are the constructible points of a given rank? This question can be addressed by recursively constructing all possible points.

Consider first the case r=0, i.e., an unmarked square. In this case, there are four identifiable points: the four corners, and four lines: the four edges. It takes no folds to identify the corners of the square; we therefore assign the four corners a rank of zero.

We can also assign a rank to a fold line as well as to a point; the rank of a line is the number of folds it takes to create the line. In an unfolded square, there are four identifiable lines, which are the four edges of the square. Since they take no folds to construct, these four lines get a rank of zero as well. Thus, a square has four distinct points and four distinct lines of rank r=0.

Now consider r=1. The possible folds for each operation are illustrated in Figure 39 for O1, O2, O3, and O5. O4, O6 and O7 do not (yet) permit the creation of any new lines.

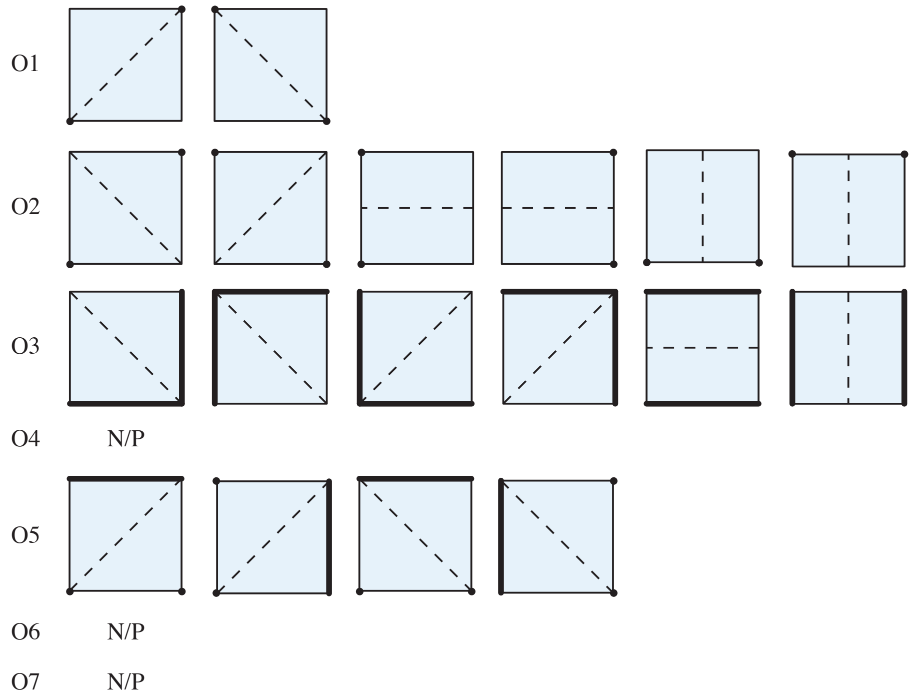


Figure 39. The constructible lines on an unmarked square using the 7 HHA operations. The points and lines involved in the construction are highlighted.

Figure 39 shows that among all HHA operations, there are four distinct new lines that can be created: the two angle bisectors and the two midlines of the square. Since each of these lines requires one fold to create, each has rank 1.

The intersections between the four new lines with each other and with the original edges of the square defines five new points: the midpoints of the sides and the very center of the square. Each new point along the edge of the square is defined by the intersection of a rank-0 edge and a rank-1 line; since they can be formed with a single fold, we therefore give them a rank of 1. The center point may be defined as the intersection of several pairs of lines, but in all combinations, both lines are rank-1; therefore, the center point is rank 2. There are now a total of 9 distinct points and 12 distinct lines. 8 of each have rank $r \leq 1$.

Now, let us consider making one more fold. With 9 distinct points, there are $9 \times 8 = 72$ possible pairs of points. Operations O1 and O2 each act on a pair of points and create a new point; thus, for this next stage of construction, we would expect to go from 9 points to 153 possible points. Similarly, O3 acts on pairs of lines, O4 on point-line combinations, and so forth. Each operation creates geometrically more lines (pairs of whose intersections define geometrically more points). Even though duplications are inevitable, the number of distinct possible lines and points increases exponentially with the number of allowed folds.

As we make more folds, the rank of the newly created folds and points can be expressed in terms of the ranks of the points and lines that are brought into alignment to create them. A new point is always defined as the intersection of exactly two lines, and its rank is given by

$$r_p = r_{l_1} + r_{l_2}. \tag{76}$$

On the other hand, a fold line can be created in several ways by various combinations of points and lines, and its rank is always increased by 1 (to account for the fold line itself):

$$r_{l_f} = 1 + \sum_{p_i} r_{p_i} + \sum_{l_j} r_{l_j}. \tag{77}$$

The number of constructible points of a given rank depends on the operations we allow for their construction. The simplest case, and one that is analytically treatable, is the double-binary method, in which we restrict ourselves to bringing pairs of points along a single edge together to make folds. It is fairly easy to show that the number of constructible point along a single edge is given by

$$N(r) = 1 + 2^r. \tag{78}$$

With somewhat more effort it can be shown that the number of constructible points of rank r located anywhere within a square using the double-binary construction is given by

$$N(r) = 1 + \left(3 + \frac{r}{2}\right)2^r, \tag{79}$$

which defines the sequence N={4, 8, 17, 37, 81…} for r={0, 1, 2…}.

If we open up the acceptable operations to include all 7 HHA operations, the combinatorics explode. Simply counting up the number of ways of combining points and lines among all possible operations gives the sequence N={4, 258, 154,800, 132,826,269…}, which grows by about a factor of 1000 with each iteration. However, only a fraction of the possible combinations are physically realizable, and those include many duplicates — identical points that can be constructed with different folding sequences. The number of *distinct* constructible points is far smaller than the combinatorial limit.

Another fly in the ointment is that knowing that a simply-constructible point is somewhere near a target point is not the same as knowing what the constructible point actually *is*. It would be nice if, given an arbitrary point (x,y), we could find a formula for the nearest constructible point of a given rank and the folding sequence for its construction.

Such a formula existed for the binary approximation of a single proportion; given a number x, the nearest constructible fraction of rank N was the N-digit binary approximation for x, and the folding sequence was encoded in the binary representation of x.

For the general case, we allow all of the HHA operations, and allow any combination of lines and points that creates a line within the square. Unfortunately, for the general case, there is no known method for efficiently finding the nearest constructible point of a given rank, and I strongly suspect that no such method exists.

Fortunately, even inefficient methods can be suitable. Since 10^6 points should suffice to provide an accuracy of around .001, it would suffice to simply construct the 10^6 or so lowest-rank marks and lines; then given a desired target point, one simply searches through them all to find the closest point. Obviously, this is not something that one does by hand; but it is quite possible for a computer.

I wrote a C++ program called *ReferenceFinder* that does just this. It takes as input the coordinates of a target reference and prints out the best folding sequences for locating that point.

In its initialization, *ReferenceFinder* constructs a database of about 300,000 distinct lines and marks of rank 6 or below, by recursively building up higher-rank marks from the lower ranks, weeding out duplicates, near-duplicates, and combinations that are not physically realizable as it goes. This fairly restrictive filtering results in a much more modest, but still impressive rate of growth in numbers of marks, which runs N={4, 8, 65, 1033, 7009, 32,469, 277,546}.

Using the 277,546 marks with rank of 6 or less, I picked 1000 random target points, found the closest constructible points, and computed statistics on the distribution of errors. The results are shown in Table 13.

Percentile	Error
10th	0.0004
20th	0.0006
50th	0.0013
80th	0.0024
90th	0.0032
95th	0.0042
99th	0.0081

Table 13. Percentile and error for sequences taken from 277,546 6-fold constructions of distinct points.

In general, an error of 0.005 — 1.2 mm out of a 25 cm square — is barely noticeable. For 97% of target points, there is a 6-fold sequence that achieves that level of error. Compare that with the binary method, which requires 18 folds to achieve the same accuracy.

The difference arises from the fact that at each stage of the construction, the number of possible distinct creases and marks is based on many possible combinations of lower-rank objects, which leads to exponential growth; the exponential scaling constant is roughly related to the number of different ways that points and lines can be combined to yield new ones.

Computer solution for efficient folding sequences is of more than academic interest. As origami designers turn to mathematical methods of designing origami, it becomes necessary to develop efficient folding sequences for reference points that are defined solely as the solution of high-order algebraic equations. Programs like *ReferenceFinder* can construct those folding sequences, which can be surprising in their efficiency. Several recent origami books [3, 35, 36] have incorporated such computer-generated folding sequences as part of the instruction of individual figures, and I anticipate that such usage will become more common in the future.

References

[1] James Brunton, "Mathematical exercises in paper folding," *Mathematics in School*, Longmans for the Mathematical Association, vol. 2, no. 4, July 1973, p. 25.

[2] Robert J. Lang, "Four Problems III," *British Origami*, no. 132, October, 1988, pp. 7–11.

[3] Robert J. Lang, "Western Pond Turtle," *Origami Design Secrets*, A K Peters, 2003.

[4] Shuzo Fujimoto and M. Nishiwaki, *Sojo Suru Origami Asobi Eno Shotai* (Invitation to creative origami playing), Asahi Culture Centre, 1982.

[5] Jeannine Mosely, private comm.

[6] Masamichi Noma, *Origami Tanteidan Newsletter*, issue 14.

[7] Koji Husimi, *Origami no kikagaku* (Origami and Geometry), Nippon Hyoronsha, Tokyo (1979).

[8] Kunihiko Kasahara and Toshie Takahama, *Origami for the Connoisseur*, Japan Publications, 1987, pp. 18–19.

[9] Kunihiko Kasahara, *Origami Omnibus*, Japan Publications, 1988, pp. 76–77.

[10] Aleksandr Ia. Khinchin, *Continued Fractions*, Dover Publications, 1997.

[11] Peter Engel, *Folding the Universe: Origami from Angelfish to Zen*, New York, Vintage Books, 1989.

[12] Jun Maekawa, "Evolution of Origami Organisms," *Symmetry: Culture and Science*, vol. 5, no. 2, 1994, pp. 167–177.

[13] Kunihiko Kasahara, *Viva! Origami*, Tokyo, Sanrio, 1983.

[14] Robert J. Lang, "Albert Joins the Fold," *New Scientist*, vol. 124, no. 1696/1697, December 23/30, pp. 38–57, 1989.

[15] Robert J. Lang, "Origami: Complexity Increasing," *Engineering & Science*, vol. 52, no. 2, pp. 16–23, 1989.

[16] David Auckly and John Cleveland, "Totally Real Origami and Impossible Paper Folding," *American Mathematical Monthly*, vol. 102, no. 3, pp. 215–226, March 1995.

[17] Thomas Hull, "A Note on 'Impossible' Paperfolding," *The American Mathematical Monthly*, vol. 103, no. 3, March, 1996.

[18] Tsune Abe, described in *British Origami*, no. 108, p. 9, 1984

[19] Koji Fusimi, "Trisection of angle by Abe," *Saiensu* supplement, October, 1980, p. 8.

[20] Jacques Justin, described in *British Origami*, no. 107, pp. 14–15, 1984.

[21] Robert Geretschläger, "Euclidean Constructions and the Geometry of Origami," *Mathematics Magazine*, vol. 68, no. 5, December, 1995, pp. 357–371.

[22] Robert Geretschläger, "Folding the Regular Triskaidekagon," presented at AMS Joint Mathematics Meeting, Baltimore, MD., January 9, 1998.

[23] Robert Geretschläger, "Folding the Regular 19-gon," presented at AMS Joint Mathematics Meeting, Baltimore, MD., January 9, 1998.

[24] Robert Geretschläger, "Solving Quartic Equations in Origami," presented at AMS Joint Mathematics Meeting, Baltimore, MD., January 9, 1998.

[25] James Pierpont, "On an undemonstrated theorem of the Disquisitiones Aritmeticae," American Mathematical Monthly Bulletin, no. 2, 1895–1896, pp. 77–83.

[26] Robert J. Lang, *The Complete Book of Origami*, New York, Dover Publications, 1988.

[27] Humiaki Huzita and Benedetto Scimemi, "The Algebra of Paper-Folding (Origami)," *Proceedings of the First International Meeting of Origami Science and Technology*, H. Huzita, ed., 1989, pp. 215–222.

[28] Humiaki Huzita, “Understanding Geometry through Origami Axioms,” *Proceedings of the First International Conference on Origami in Education and Therapy (COET91),* J. Smith ed., British Origami Society, 1992, pp. 37–70.

[29] Thomas Hull, “Geometric Constructions via Origami,” *Proceedings of the Second International Conference on Origami in Education and Therapy (COET95*), V’Ann Cornelius, ed., Origami USA, 1995, pp. 31–38.

[30] Peter Messer, “Problem 1054,” *Crux Mathematigorum*, vol. 12, no. 10, December, 1986.

[31] Thomas Hull, http://web.merrimack.edu/hullt/geoconst.html, 2003.

[32] Koshiro Hatori, http://www.jade.dti.ne.jp/~hatori/library/conste.html, 2003.

[33] http://mathworld.wolfram.com/Ophiuride.html, 2003.

[34] http://mathworld.wolfram.com/CissoidofDiocles.html, 2003.

[35] Robert J. Lang, *Origami Insects II*, Gallery Origami House, Tokyo, 2003 [in press].

[36] John Montroll, *A Plethora of Polyhedra in Origami*, Dover Publications, 2002.